A GENERIC SEMANTIC GRAMMAR FOR QUANTITATIVE TEXT ANALYSIS: APPLICATIONS TO EAST AND WEST BERLIN RADIO NEWS CONTENT FROM 1979

Carl W. Roberts*

In a semantic text analysis the researcher begins by creating one of two types of semantic grammars, each of which provides one or more templates that specify the ways concepts (or more general themes) may be related. On the one hand, a phenomenal semantic grammar can be created to extract phenomenon-related information from a text population (e.g., “Among the population’s grievances [the phenomenon of interest in this case], which were ones for the abolition of taxes?”). On the other hand, a generic semantic grammar may be developed to yield data about the text population itself (e.g., “Among all clauses in the text population, how many were grievances for the abolition of taxes?”). This paper describes a generic semantic grammar that can be used to encode themes and theme relations in every clause within randomly sampled texts. Unlike the surface-grammatical relations mapped by syntax grammars, the theme relations allowed in this grammar only permit unambiguous encoding according to the meanings that clauses were intended to convey within their social context. An application of the grammar provides a concrete illustration of its research potential.

This paper was presented at the American Sociological Association meeting in New York City, August 1996. I am particularly indebted to Albert Baker, whose careful eye added greatly to the rigor and at times to the substance of the formalism developed in this paper, and to Beth Lencowski, whose long hours of coding and analysis were of immeasurable assistance. E-mail: carlos@iastate.edu

*Iowa State University
Semantic text analysis is a quantitative text analysis method in which not only themes (or classes of concepts) but also grammatical relations among themes are encoded. The method involves a three-step encoding process. First, the researcher isolates a population of texts germane to the phenomenon under investigation. Second, a semantic grammar must be acquired that specifies the relations that may be encoded among themes in the texts. Finally, the texts’ themes are encoded according to the relations specified in the semantic grammar. Encoded interrelations may then be used as indicators of various characteristics of the phenomenon under study.

Consider, for example, the text population of *cahiers de doléances* of 1789 that Markoff, Shapiro, and their colleagues have been analyzing since the mid-1970s (cf. Markoff 1988; Markoff, Shapiro, and Weitman 1974; Shapiro and Markoff, forthcoming). The *cahiers de doléances* are the documents produced by more than 40,000 corporate and territorial entities (e.g., craft guilds, parishes, towns, etc.) in the course of the king’s convocation of an Estates-General—documents written as if in response to the open-ended query, “What are your grievances, and what do you propose should be done about them?” The documents were used to generate data on grievances, making grievances the researchers’ phenomena of interest and, for the purposes of constructing a data base for these grievances, effectively making “the grievance” the unit of analysis.

In essence, each grievance was encoded according to two syntactic components. (Or, if you prefer, a template with two interrelated fields was used in encoding each grievance.) First, there was the institution or problem area (i.e., the thing being grieved about). These grievances might be about the government, the economy, religion, the constitution, and so on. The second syntactic component of a grievance was the action demanded. These actions were encoded as demands to reestablish, to abolish, to simplify, to modify, to improve, and so on. In brief, the researchers designed a two-place Verb-Object (V-O) semantic grammar for the phenomenon, grievance. The semantic grammar’s application to a sample of texts from the *cahiers de doléances* yielded data that have been used to make infer-

---

1 Throughout this paper I use “the phenomenon under investigation,” “the phenomenon under study,” “the phenomenon of interest,” and “the unit of analysis” interchangeably. All refer to that phenomenon (as defined in the researcher’s theory), sampled instances of which correspond to distinct rows in the researcher’s data matrix, and measures of which are listed in the columns of this matrix.

2 For each action (verb) toward a problem area (object), it was uniformly the king and his representatives (subject) who should act. Thus a more detailed Subject-Verb-Object syntax was unnecessary. This discussion presents only a simplified ver-
ences about public opinion on a variety of topics just prior to the French Revolution.

A second illustration can be found with Franzosi’s (1994) methodological writings on labor disputes. Applying a more complex semantic grammar to newspaper articles on Italian labor unrest, labor disputes (the phenomena of interest) are conceptualized as clusters of actors’ actions toward each other (i.e., of Subject-Verb-Object [S-V-O] tuples). Like Markoff and Shapiro, Franzosi has used texts as a source of historical data. Yet whereas the former researchers commonly found more than one grievance per document (i.e., multiple units of analysis per text block), Franzosi had numerous S-V-O tuples per newspaper article and commonly multiple newspaper articles per labor dispute (i.e., multiple text blocks per unit of analysis). Thus, for Franzosi, the generation of dispute-specific indicators of whether one type of actor acted in a specific way toward another, requires a search for this information among all S-V-O tuples associated with each labor dispute.

1. SEMANTIC GRAMMARS HIGHLIGHT STRUCTURES WITHIN TEXT POPULATIONS

The preceding illustrations demonstrate that in applying a semantic grammar to a sample of texts, the researcher assumes that the texts are structured according to the semantic grammar and that the phenomenon of interest is related to this structure in a specific way. In Franzosi’s work, for example, newspaper articles are presumed to contain information on actors’ actions toward each other, and labor disputes are portrayed as consisting of clusters of such actions. Thus every semantic text analysis must begin by isolating a population of texts that exhibit the structure assumed in the research at hand.

On the basis of similar observations, Griemans (1984 [1966]) cautioned linguists to assemble a representative, exhaustive, and homogeneous corpus of texts prior to beginning analysis. Even more strongly put, Hallow (1978:32) argued that every utterance (or speech act) must be unders-

5This warning is quite different from that of the statistician who notes that (other things equal) analyses of data from homogeneous populations will afford statistics with small standard errors. Griemans’ point is that one must ensure the relevance of one’s corpus to one’s semantic model: “[A] model can be described only if it is already contained implicitly in the discursive manifestation of a semantic microuniverse” (Griemans 1984[1966]: 163).
stood according to its “context of situation.” Moreover, these situational contexts impose structural constraints on what statements are socially appropriate (cf. Lakoff and Johnson 1980:179 on the fit between statement and situation).

Linguists developed semantic grammars (or functional grammars; cf. Halliday 1978, 1994; Propp 1968 [1928]) as a strategy for describing text structure. Application of the strategy begins by identifying speech acts according to how they function within the genre of texts under analysis (e.g., as stating a situation, explaining a problem, responding to the problem, evaluating the response). The genre (e.g., of texts with problem-solution structure; cf. Hoey 1994) is then characterized according to the sequence of functional forms common among its texts. The grammars underlying such sequences have been variously referred to as narrative grammars (Griemass 1984 [1966]), text grammars (van Dijk 1972), and story grammars (Rumelhart 1975).

In fields other than linguistics, social scientists have tended to be less interested in the form than in the content of texts’ grammars. In analyses of text sequence, this preference of content to form holds among social scientists who are not primarily linguists but who have orientations both qualitative (Abell 1987; Heise 1991) and quantitative (Namenwirth 1973; Schrot 1991). In contrast to linguists’ objective to reveal text structure as a sequence of distinct forms, in a semantic text analysis one generally presumes a single semantic form with varying content. Whereas linguists typically revise their semantic grammars to fit a relevant corpus, semantic text analysts use fixed semantic grammars to highlight relevant text.

1.1. Phenomenal Versus Generic Highlighting of Text Structure

In a semantic text analysis, the researcher encodes only those parts of the text that fit into the syntactic components of the semantic grammar being applied. For example, a preliminary statement in a document from the cahiers de doléances that “The members of this guild have always honorably served our King” would quickly be recognized by the coder as not conveying a grievance and would accordingly be ignored. Likewise, Franszois’s semantic grammar does not lend itself readily to evaluative statements (e.g., a reporter’s aside that the police had acted inappropriately during a strike). In such cases, the researcher does not experience the linguist’s concern that the grammar might not fit the texts but instead notes
that the semantic grammar highlights only text structures that fit the grammar (and presumably only those text structures that are relevant to the phenomenon under investigation).

The semantic grammar’s highlighting role has methodological advantages when one’s research objective is to analyze variations among the aspects of a well-defined phenomenon (e.g., the grievance or the labor dispute). By restricting encoding to text segments with relevance to the phenomenon of interest, such a phenomenal semantic grammar will save both time and expense. Yet the highlighting role works poorly when one’s research objective requires a more generic semantic grammar for investigating the predominance of theme relations within randomly sampled text blocks that themselves do not correspond to specific, highly structured phenomena. In the former case, the researcher encodes only themes that are related according to a semantic grammar; in the latter case, the researcher encodes all data and investigates the conditions under which specific theme relations occur. Research questions posed, for example, in cultural indicators and media research lend themselves more readily to applications of generic than phenomenal semantic grammars: Is government depicted in totalitarian states’ news media as the semantic subject (e.g., acting in the people’s interests), but in democratic states’ news media as the semantic object (e.g., affected by the people’s wishes)? On prime-time television, do men utter fewer degrading self-references than women? Here the researcher is not interested in examining only relevant phenomena (be they government depictions or self-references) within a text population but rather in determining whether certain types of content are prevalent relative to all content in a population of texts.

1.2. Surface Grammar’s Highlighting of Ambiguous Text Structure

A generic semantic grammar is required to encode interrelations among themes within a domain of relatively unstructured texts (e.g., general content from prime-time television). The most obvious candidate for this semantic grammar is a S-V-O grammar to be applied to each clause in samples of such text.\(^4\) In drawing inferences about the predominance of theme

\(^4\)The reader familiar with literature on semantic grammars will recognize a contradiction in terms when I refer to a clause’s surface (or syntax) grammar as a type of semantic grammar. Semantic grammars were developed to map statements’ unique meaning-as-intended and in so doing to avoid mapping superficial grammatical relations that could have many intended meanings. I return to this issue in the next section.
relations, the researcher could then treat either the clause or (collapsing data across clauses) the text block as the unit of analysis.

Both Gottschalk (1968, 1995) and Schrodt (1993) developed semantic text analysis methods that incorporate precisely this type of generic S-V-O grammar. Taking advantage of the fact that their methods encode surface-grammatical relations among themes, each has incorporated a parser into his software that identifies which of each clause’s themes functions as which of the three syntactic components, subject, verb, and object. The Gottschalk-Gleser content analysis system outputs aggregate scores on individuals’ psychological states (e.g., anxiety, hostility, etc.); Schrodt’s KEDS program outputs unaggregated S-V-O tuples.5

Yet linguists have long argued that texts’ intended meanings are not captured solely by their surface grammatical relations. Indeed, it is precisely this realization that led them to develop semantic grammars and to distinguish these grammars from more semantically ambiguous syntax grammars. For example, *He was abandoned* might either refer to a state of affairs (i.e., he was alone) or to a process (i.e., others had abandoned him). *She is a doctor* might be intended as descriptive (implying that she helps sick people) or judgmental (indicating that she is an achiever). In short, semantic grammars require the coder to take clauses’ social context into account; syntax grammars do not.

Nonetheless, researchers may have legitimate reasons to base their semantic text analyses on syntax grammar. People often betray their mental states in the ways they phrase their discourse, making the words’ surface phrasing more relevant than their intended meanings in making a psychological diagnosis (cf. Gottschalk 1995). On the other hand, if one’s analyses are of a sufficiently structured domain of texts (e.g., Reuters news service articles on international conflict), theme relations may follow sufficiently fixed formulae that their surface relations are nearly always unambiguous (cf. Schrodt 1993).

Venturing beyond such cases, one finds syntax grammars fundamentally inadequate for addressing research questions that call for the analysis of words’ intended meanings—a central tenet among functional

---

5 Both Gottschalk and Schrodt take an instrumental approach to text analysis. According to Shapiro (1997) instrumental text analyses treat the text as symptomatic of the phenomenon of interest (for Gottschalk, the individual; for Schrodt, the political event). Because coders are not required to divine clauses’ intended meanings, the encoding of theme-relations can be largely automated with the help of parsing software. All other semantic text analysis methods discussed in this paper are what Shapiro refers to as representational. That is, they are text analysis methods in which texts are encoded according to their sources’ intended meanings.
linguists such as Halliday (1978:192) and Winter (1994:49). When clauses are embedded in texts that are unstructured enough for identical speech acts to serve different discursive functions (i.e., to have different intended meanings), surface grammatical relations cannot differentiate among these functions and thus cannot specify a clause’s intended meaning(s).\textsuperscript{6} Inferences about such text populations call for a generic semantic grammar that allows clauses to be encoded according to discursive function.

2. A GENERIC GRAMMAR FOR SEMANTIC TEXT ANALYSIS

It was Gottlob Frege (1959 [1884]), who first noted that the sentence, \textit{x acts}, makes two assertions: “there is an x” and “x acts.” That is, sentences of this form simultaneously describe a state of affairs and a process. This dual form is commonly expressed with the following notation:

\[(x)f(x)\]

The first element in the form (namely, “(x)”\textsuperscript{6}) may be expressed as “There is an x.” The second element (namely, “f(x)”\textsuperscript{6}) is read as “f(\cdot) is a predicate involving x.” If f(\cdot) predicates acting, f(x) predicates x as acting. As I now formally argue, this rendering is ambiguous as a functional discourse form and as such cannot (without modification) serve as a basis for a generic semantic grammar. Researchers interested in encoding texts according to their intended meanings will have little use for grammars that afford semantically ambiguous mappings from text to meaning. My argument begins with a demonstration that a grammar’s functional forms are semantically ambiguous (and thus of little value to these researchers) if they do not have unique semantic opposites.

2.1. Semantic Opposition in Ordinary Discourse

This subsection provides the theoretical basis for a generic semantic grammar composed of four unambiguous functional forms (i.e., functional forms having unique semantic opposites). In it, two functional forms for descrip-

\textsuperscript{6}The assumption here is not that speech acts can serve only a single intended function. On the contrary, speech acts are often intended to affect others in ambivalent ways. The assumption is instead that in natural language each clause is uttered to function in an enumerable number of ways. Thus a coder could in principle apply a semantic grammar to encode the same clause numerous times, each time according to a distinct function that the clause was intended to serve.
tive speech are developed as components in a model of speech acts’ intended meanings. The argument here is that in ordinary discourse a speech act’s meaning consists of an unintentional, taken-for-granted component plus an intentional, asserted component. The unintentional component is neither denied nor asserted by the source but is simply assumed to be common knowledge. The intentional component’s meaning is asserted (and, equivalently, its semantic opposite is denied) in the speech act. The ensuing discussion reveals a structure of linguistic ambiguity within ordinary discourse by showing that descriptive utterances admit of precisely two semantic opposites. This motivates a more formal specification in the next subsection of two unambiguous functional forms for descriptive speech acts. These are then supplemented by another two for judgmental speech acts.

A semantic opposite differs from a logical opposite in that it is the negation of the intended meaning, not the literal formal meaning, of a speech act. For example, consider the sentence, *Jerry went to the store.* Applying the \((x)f(x)\) functional form, and setting \(x = \text{“Jerry”}\) and \(f(\cdot) = \text{“went to the store”}\) yields “There is Jerry and Jerry went to the store.” In formal Aristotelian logic, the sentence is rendered as follows:

\[
\exists (\text{Phenomenon } x) \exists (\text{Process } p) \left[p(x) \land x = \text{“Jerry”}\right]
\]

\[
\land p = \text{“went to the store”}\]

This statement reads, “There exists a phenomenon, \(x\), such that (sic) there exists a process, \(p\), such that \(p\) is predicated of \(x\), \(x\) is ‘Jerry’, and \(p\) is ‘went to the store’.” Note that this statement has the following as its unique logical opposite:

\[
\forall (\text{Phenomenon } x) \forall (\text{Process } p) \left[\neg p(x) \lor x \neq \text{“Jerry”}\right]
\]

\[
\lor p \neq \text{“went to the store”}\]

(I.e., for all phenomena, \(x\), such that for all processes, \(p\), \(p\) is not predicated of \(x\) or \(x\) is not “Jerry” or \(p\) is not “went to the store.”) However, in analyses of ordinary discourse the semantic opposite of the sentence’s intended meaning has a much narrower scope.

In ordinary discourse, the efficient functioning of natural language requires that both source and audience take much of the original Aristotelian expression’s content for granted. That is to say, most elements of the expression will be assumed semantically invariant and thus superfluous to its intended meaning. There are four such elements.

- \(\exists (\text{Phenomenon } x) [x = \cdot]\) — The speech act mentions physical and symbolic phenomena that are true to the audience’s experiences. Thus in
ordinary discourse a loyal subject does not intend to communicate that the emperor's clothes exist when making references to their elegance. An emperor without clothes could not occur.

- $\exists (\text{Process } p) [p = \cdot]$—The speech act relates phenomena in ways that are comprehensible to an audience fluent in the language of discourse. Thus in ordinary discourse the source does not intend to communicate the existence of processes such as "going to the store." Discourse will be sidetracked when such processes require definition.

- $p(x)$—The source genuinely intends to communicate a process predicated on a phenomenon. That is, in ordinary discourse the source does not intend to communicate that a phenomenon and a process are being linked. Were the audience to begin attending to the appropriateness of the $p$ on $x$ link, the source's credibility could be called into question. Accordingly, the statement, "We Grecians offer the citizens of Troy a great wooden horse as a gift," was not understood by Laocoön as intended to describe an event but to link "Grecians" with "gift giving" in the minds of his fellow Trojans.

- $\neg (x \neq \text{"Jerry" } \land p \neq \text{"went to the store"})$—The source intends to communicate relevant information. That is, in ordinary discourse the source does not intend communications that have uninformative semantic opposites. It is for this reason that the following could not be the semantic opposite of "Jodi ran away with a circus":

$$\exists (\text{Phenomenon } x) \exists (\text{Process } p) [p(x) \land x \neq \text{"Jodi"} \land p \neq \text{"ran away with a circus"}]$$

If this were the semantic opposite of the source's intended meaning, "Jodi ran away with a circus" would comprise a denial that "Something 'other than Jodi' did something, which was 'something other than' running away with a circus"—a remarkably uninformative statement. When the audience discovers that such an uninformative denial was intended (e.g., as the source continues, "But the police made her bring it back"), a humorous departure from ordinary discourse results.

Thus, if the audience assumes the truth, comprehensibility, credibility, and relevance of the source's speech acts, the sentence, Jerry went to the store, has exactly two semantic opposites: 7

7The audience's assumptions of truth, comprehensibility, credibility, and relevance have direct parallels with Habermas's (1979:1-68) discussions of validity claims that in ordinary communication are respectively true, comprehensible, truthful, and right.
\[ \exists (\text{Phenomenon } x) \exists (\text{Process } p) \left[ p(x) \land x \neq \text{ "Jerry"} \land p = \text{ "went to the store"} \right] \]

and

\[ \exists (\text{Phenomenon } x) \exists (\text{Process } p) \left[ p(x) \land x = \text{ "Jerry"} \land p \neq \text{ "went to the store"} \right] \]

Once the domain of a generic semantic grammar is restricted to ordinary discourse (i.e., to speech acts that the audience assumes true, comprehensible, credible, and relevant), the simpler \((x)f(x)\) notation can be substituted for expressions of Aristotelian logic. Accordingly, the functional forms of the two just-mentioned semantic opposites are as follows:

\[(x)f(\neg x) \quad \text{and} \quad (x)\neg f(x)\]

When applied to the sentence, *Jerry went to the store*, the first form’s transformation can be read as, “Something ‘other than Jerry’ went to the store,” whereas the second transformation generates the semantic opposite, “Jerry did ‘something other than’ go to the store.”

It is because of these dual semantic opposites that all sentences fitting the \((x)f(x)\) form are ambiguous. Differently put, the intended meaning of the sentence about Jerry depends on whether its function was to answer the question, “Who or what went to the store?” or “What did Jerry do?” In the former case, the sentence functions to convey a description of a state of affairs; in the latter case, it functions to convey a description of a process.

### 2.2. Four Unambiguous Functional Forms

What this rather lengthy illustration suggests is that a semantic grammar cannot yield unambiguous encoding of texts unless all its functional forms have unique semantic opposites. Let us assume that the discursive function of the sentence, *Jerry went to the store*, was to identify who or what was

---

8 Of course, the sentence might instead have been intended to answer the question, “Where did Jerry go?” Or it might have been intended to answer more than one of these questions. The reader is referred to note 6 regarding this latter point. Regarding the former point: In addressing the where-did-Jerry-go question, the sentence would function to convey a description of a state of affairs. Because this function is one of four being suggested in this paper as the basis for a generic semantic grammar, it is a case perfectly consistent with the argument at hand.
storebound. The following functional form provides a syntax for such descriptions of states of affairs.\textsuperscript{9}

\[(x) a(x) \quad \text{with semantic opposite} \quad (x) a(\sim x)\]

A fit between sentence and form might be read as, “There was (a storebound) Jerry,” with semantic opposite, “There was no (storebound) Jerry.”\textsuperscript{10} If in a different context the discursive function of the same sentence were to convey a description of a process, the appropriate functional form would be:\textsuperscript{11}

\[(x)^i p(x) \quad \text{with semantic opposite} \quad (x)^i \sim p(x), \quad i = 0, 1\]

The i-superscript is introduced here to acknowledge the optional role of the semantic subject in passive voice. Accordingly, if the semantic subject, Jerry, had not been named in the sentence being encoded, its rendering might be read as, “The store was gone to,” with semantic opposite, “Something other than going to the store happened.”

Ambiguity is also present when a clause functions to convey a positive (or negative) judgment of a description’s referent. For example, the sentence, Chris makes charitable contributions, may function to convey not only a description but also a positive evaluation of Chris (or possibly a positive evaluation of “making charitable contributions,” if Chris is one of

\textsuperscript{9}More formal renderings of \((x) a(x)\) and \((x) a(\sim x)\) are respectively \(\exists(\text{Phenomenon } x) \exists(\text{Attribute } a) [a(x) \land x = \text{“Jerry” } \land a = \text{“storebound”}]\) and \(\exists(\text{Phenomenon } x) \exists(\text{Attribute } a) [a(x) \land x \neq \text{“Jerry” } \land a = \text{“storebound”}]\).

\textsuperscript{10}Linguists commonly represent “being” and “becoming” as functions with two arguments, such that the content of one argument can be represented as being or becoming that of the other. Linguistic content analysis (Roberts 1989) uses a functional form that renders descriptions of states of affairs as one of four, two-place predicates (”is an instance of,” “becomes,” “resembles,” or “symbolizes”). When a two-place predicate is used, it is the posterior, not the anterior, argument that is negated in the clause’s opposite. For example, it is semantically identical to assert that “Jerry went to the store” (with opposite “Jerry did ‘something other than’ go to the store”) or that “Jerry was an instance of ‘a storebound entity’” (with opposite “Jerry was not an instance of ‘a storebound entity’”). In both cases, the sentence presumes Jerry’s existence while functioning to convey a description of his storebound activity. If one encodes “Jerry” as the posterior argument in the two-place predicate, its rendering (i.e., “The storebound entity was Jerry” with opposite “The storebound entity was not Jerry”) not only is semantically distinct from both other renderings but also functions to convey a description of a state of affairs and, identically, the answer to “Who or what went to the store?”

\textsuperscript{11}Formal renderings of \((x) p(x)\) and \((x) \sim p(x)\) are respectively \(\exists(\text{Phenomenon } x) \exists(\text{Process } p) [p(x) \land x = \text{“Jerry” } \land p = \text{“went to the store”}]\) and \(\exists(\text{Phenomenon } x) \exists(\text{Process } p) [p(x) \land x = \text{“Jerry” } \land p \neq \text{“went to the store”}]\).
the speaker’s heroes). For such cases, a functional form can be introduced that renders a positive judgment of a state of affairs as follows:  \[ (x) a(x) Q_a \] with semantic opposite \[ (x) a(x) \overline{Q}_a \]

In the former expression \( x \)'s attribute, \( a \), is assigned the positive qualifier, \( Q_a \). In the latter expression the same attribute is assigned the opposite, negative qualifier, \( \overline{Q}_a \). Thus, if the above sentence was solely intended to convey a positive judgment of Chris, its unique semantic opposite could be rendered as “Chris is a bad person (presumably by virtue of the egregious nature of those who contribute to charities).”

A formal representation of a positive judgment of a process can be rendered as follows:  \[ (x)^i p(x) Q_p \] with semantic opposite \[ (x)^i p(x) \overline{Q}_p, \quad i = 0 \]

Here the \( i \)-superscript acknowledges that a state of affairs need not be explicit in such speech acts. (For example, one may assert, “Dancing on Sunday is immoral.” without naming a particular dancer.) In the former expression, the process \( p(x) \) is assigned a positive qualifier, \( Q_p \). In the latter expression, the same process is assigned the opposite, negative qualifier, \( \overline{Q}_p \). Accordingly, if the sole intention of the speech act about Chris was to convey a positive judgment of his behavior, its unique semantic opposite could be rendered as “Chris’s making of charitable contributions is immoral.”

Note that in these last two judgmental functional forms, all elements except the quantifiers are semantically invariant. Differently put, in ordinary discourse the extent to which a speech act is intended to positively judge a process (or state of affairs) is the extent to which it is intended to deny a negative judgment of the same process (or state). If a speech act were intended to convey, “Chris is a good person,” its intention would not be to deny that someone other than Chris is a good person but

---

12 Keeping with the earlier illustration, more formal renderings of \( (x)a(x)Q_a \) and \( (x)a(x)\overline{Q}_a \) are respectively \( \exists(\text{Phenomenon \( x \)}) \exists(\text{Attribute \( a \)}) \exists(\text{EvalFunction \( e_A \)}) \quad [a(x) \wedge x = "\text{Jerry}" \wedge a = "\text{storebound}" \wedge e_A(a)] \) and \( \exists(\text{Phenomenon \( x \)}) \exists(\text{Attribute \( a \)}) \exists(\text{EvalFunction \( e_A \)}) \quad [a(x) \wedge x = "\text{Jerry}" \wedge a = "\text{storebound}" \wedge \overline{e_A(a)}] \). Here the function, \( e_A \), assigns a positive (and \( \overline{e_A} \) a negative) judgment to the attribute, \( a \), contained in the set \( A \).

13 More formal renderings of \( (x)^i f(x)Q_p \) and \( (x)^i f(x)\overline{Q}_p \) are respectively \( \exists(\text{Phenomenon \( x \)}) \exists(\text{Process \( p \)}) \exists(\text{JustFunction \( j_f \)}) \quad [p(x) \wedge x = "\text{Jerry}" \wedge p = "\text{went to the store}" \wedge j_f(p)] \) and \( \exists(\text{Phenomenon \( x \)}) \exists(\text{Process \( p \)}) \exists(\text{JustFunction \( j_f \)}) \quad [p(x) \wedge x = "\text{Jerry}" \wedge p = "\text{went to the store}" \wedge j_f(p)] \). Here the function, \( j_f \), assigns a positive (and \( j_f \) a negative) judgment to the process, \( p \), contained in the set \( P \).
rather to deny that Chris is a bad person. Likewise, if a speech act were intended to convey, "Dancing on Sunday is immoral," its intention is not to deny that activities other than Sunday dancing are immoral but rather to deny that Sunday dancing is moral.

In summary, when applying a generic semantic grammar to relatively unstructured texts, the coder’s task is not one of identifying themes’ surface grammatical relations but one of identifying each theme’s role within the functional form(s) appropriate to its clause of origin. During the coding process such identifications can, of course, only be made after selecting the appropriate functional form, a selection that requires the coder to look beyond the clause. That is, the selection of functional forms requires that the coder understands both the source’s intentions and the social context within which the clause appeared.

2.3. Criteria for Selection Among Functional Forms

Understanding intentions requires coder intuition about the mental process that (by having uttered a clause) the source attempted to initiate in an audience. The preceding discussion hints at four such intended mental processes: the recognition of a state of affairs, a perception or imagination of a process, a positive or negative evaluation of a state of affairs, and a positive or negative judgment (justification?) of a process. Accordingly,

14 Of course, "Chris is a good person" is a phrase that might be used in an attempt to communicate that someone other than Chris is a bad person (and "Dancing on Sunday is immoral" might be used to suggest that activities other than Sunday dancing are moral). In this, as in other paragraphs, phrases between quotation marks represent intended meanings, not direct quotations.

15 In different contexts others have also held the premise that speech is motivated by a desire to affect one’s audience’s mental state: "A good writer or raconteur perhaps has the power to initiate a process very similar to the one that occurs when we are actually perceiving (or imagining) events instead of merely reading or hearing about them" (Johnson-Laird 1970:270). "The production of sound for the purpose of attracting attention is language, once we have reason to assert that ‘attracting attention’ is a meaning that fits in with the functional potential of language" (Halliday, 1978:19).

16 Rough parallels exist between these four types of intention and the classes some linguists have developed to capture the illocutionary (i.e., intended) force of speech acts (Austin 1975; Searle and Vanderveken 1985) or the structurally apparent features of an aspect grammar (Dowty 1979). For example, among Austin’s classes of performative utterances, "exercitives" and "behabitives" respectively function to convey descriptions and judgments of processes. (His "expositives" and "verdictives" have much rougher correspondence to the respective functions of conveying descriptions and judgments of states of affairs.) Dowty’s "statives," "activities," "achievements," and "accomplishments" have clear parallels to the respective functions of conveying descriptions of states of affairs, descriptions of processes, judgments of states of affairs, and judgments of processes. Turning to functional linguistics, one
any matching of clause to functional form(s) requires that the coder weigh the relative plausibility of each of these four manners in which the clause may have been intended. By providing a framework for them to make such “judgments of subjective plausibility,” the four functional forms provide coders with structure in their strivings to map sources’ inherently ambiguous meanings into unambiguous code.

This structuring of the coding process has interesting parallels in Weber’s (1973 [1906]:290) recommendation that in their quest for the inherently elusive causes of past events, historical sociologists should guide their analyses by comparing numerous “judgments of objective possibility” (objektiver Möglichkeiturteile). The idea here is that a causal understanding of history requires more than knowing the events that took place; it requires judgments about whether events would have taken place in the presence of various counterfactuals (i.e., “contrary-to-fact historical alternative(s) . . . conceptually and empirically quite close to the ‘real past’”; Griffin 1993:1101-102).17 In his ETHNO program, Heise (1988-1991) provides a structure for such counterfactual inquiry by requiring the user to identify causal links among the actions within a chronology. Similar analytic rigor is gained as coders (possibly under software guidance) are required to identify the functional form(s) judged most appropriate to the clauses in one’s texts.

While weighing the relative plausibility of functional forms, the coder may judge the source not to have been sincere but instead to have attempted communication of irony, hyperbole, or understatement.18 Interestingly, the coder’s recognition of such discourse styles can only aid the coding process: Cases of hyperbole and understatement will usually suggest a clause’s judgmental rather than descriptive intent. For example, We have enough food to feed an army may be used to convey a positive or negative evaluation of the amount of food on hand. Ironic, possibly sarcastic, speech acts must be encoded according to the semantic opposite of the functional form that would apply, were they to have been uttered in

might also draw rough parallels between Winter’s (1994) two sets of basic text structures, “situation and evaluation” and “hypothetical and real” and, respectively, my “recognition and evaluation” and, less apparently, “justification and perception.”

17I am indebted to an anonymous reviewer for bringing the literature on counterfactuals to my attention.

18A source’s intention to deliberately deceive its audience is unrelated to the coder’s selection among functional forms. The semantic grammar proposed here only captures the meanings that sources intend to convey, not the pragmatics of why they opt to convey these meanings. Nonetheless, there is no reason why the coding of particular clauses might not be supplemented with identifiers that indicate the coder’s suspicions regarding the genuineness of sources’ intentions.
sincerity. For example, when said of a sputtering jalopy, the utterance, *It runs ‘beautifully,’ doesn’t it,* should be encoded as a negative judgment of the jalopy’s running (a process).

Finally, applying a semantic grammar requires that speech acts are understood in terms of their respective situational contexts. As always, these situational contexts are those within which the sources believed their speech acts would be interpreted. Invalid code will almost surely result when coders are insufficiently familiar with these contexts. However, even if the coder attained such contextual integrity in apprehending the source’s intended meanings, these meanings may not have been those understood by their intended audiences. Little substantive import is likely to be found in an analysis of texts generated by sources who themselves were unable reliably to predict audience reactions (leaving their speech, one might say, little more than sound and fury). Moreover, encoding will be futile if the source’s intended meanings (e.g., regarding preferred cookie recipes) are not relevant to the researcher’s purposes (e.g., studying political attitudes).19 Thus the four functional forms are correctly applied within the source’s situational context—a context hopefully familiar to both coder and audience, and relevant to both text population and research objectives.

### 2.4. The Grammar

This generic semantic grammar is a model of text (T), according to which text is a sequence of one or more clauses (C) separated by markers (&) that indicate subordination, coordination, and sentence boundaries:

\[ T = C(&C)^n, \quad n \geq 0 \]

Each clause in this sequence can be represented within its situational context (SC) according to one or more of the four unambiguous functional forms just introduced. Differently put, intention (I) is a function that maps (\(\rightarrow\)) clause-context pairs into a multiset comprised of subsets having one or more occurrences among recognition (R), perception (P), evaluation

---

19 To determine the relevance of texts to a semantic grammar, one may begin by encoding a small representative sample from one’s text population. If a phenomenological semantic grammar appears only to highlight a small proportion of the population, it follows that the unhighlighted proportion is not relevant to one’s research. Of course, any text with identifiable clauses can be encoded using a generic semantic grammar. Yet, if application of a generic semantic grammar (or, for that matter, of software for representational thematic text analysis; cf. Popping 1997) yields no common vocabulary among large segments of the population, these segments are likely to have little basis of comparison. In such cases, it is reasonable to consider whether one may have selected an inappropriate text population.
(E), and justification (J):

\[ I(C_i, SC_{i-1}) \rightarrow MS_{[R,PE,J]}, \]

where \( MS \neq \{ \} \)

and \( R = (x)a(x) \) with semantic opposite \( (x)a(\neg x) \)

\( P = (x)p(x) \) with semantic opposite \( (x)\neg p(x) \)

\( E = (x)a(x)Q_a \) with semantic opposite \( (x)a(x)\neg Q_a \)

\( J = (x)p(x)Q_p \) with semantic opposite \( (x)p(x)\neg Q_p \)

The restriction that this intention function have a single situational context as an argument ensures the common ground (i.e., common assumptions made by the source regarding truth, comprehensibility, credibility, relevance, and evaluation and justification criteria) needed for a one-to-one mapping from clauses’ surface representations to their appropriate functional forms.

Finally, situational contexts are continually updated by virtue of successive clauses’ intended meanings. If \( SC_0 \) is the situational context at the beginning of one’s text and \( SC_i \) is the situational context after the \( i \)th clause, situation (S) can now be introduced as a function that takes the current clause’s intention into account in transforming one situational context to the next:

\[ S(SC_{i-1}, I(C_i, SC_{i-1})) = SC_i \]

To the extent that situational contexts (and rules for transformations among them) lack formal definition, coders’ intuitions regarding these contexts and transformations will be necessary when this generic semantic grammar is used in encoding texts.\(^{20}\)

\(^{20}\)In practice, applications of this generic semantic grammar will usually call for greater complexity than is provided here. For example, specific types of questions, modality, and clause coordination and subordination may be required. Speaker- and audience-identifiers could be used in encoding direct quotations. Moreover, the grammar contains only Boolean valences and qualifiers, leaving it unable to capture their dimensions of potency and type, as well as that of positive or negative sign. I have refrained from formally specifying these components (and thus from adding to the grammar’s complexity), because it would only serve to divert attention from the paper’s central contribution to the content analysis literature: The novelty of this paper is its formal specification and illustration of a linguistically grounded semantic grammar for encoding arbitrary clauses according to their sources’ intended meanings. In lieu of a formal specification of the additional components just mentioned, these components are given lengthy illustration in the following section (particularly in conjunction with note 27 and Tables 2 and 3). More detail on qualifiers can be found in Section 3.1. The
3. AN APPLICATION OF THE SEMANTIC GRAMMAR

For illustrative purposes, let us consider the population of all clauses uttered during East and West Berlin radio news coverage of two brief military conflicts during the spring of 1979. One of these conflicts, the concluding battles of the Sandinista Revolution (SR) in Nicaragua, was an event widely acclaimed by East Germany and other communist countries, but less enthusiastically received by the West. The other conflict, the Chinese invasion (CI) of Vietnam, was vigorously denounced in East Germany and other Soviet-aligned communist countries, and received more pedestrian treatment in the West. The following is an illustration of one of these news broadcasts:21

Despite their withdrawal-notifications on Monday, Peking’s aggression-troops still provoke new fighting. Thus, the north Vietnamese provincial capital, Lan Shon, was subjected to the Chinese invaders’ sporadic artillery fire. According to agency reports, shells struck the main streets flagrantly anew. The train station and the provincial hospital are said to be completely destroyed. As the station, ‘Voice of Vietnam’, reported, as yet there is still no indication of a real withdrawal of the Peking troops. At a few locations the Chinese were said to have removed only certain ranks. In other regions they were said still to hold their positions. During their movements the invaders continued to commit barbaric crimes. According to statements from the news agency, VNA, China does not even resist attacking foreign journalists. Sharpshooters are said to have opened fire on a group of correspondents in Lan Shon, whereby they murdered Isau Takano, a reporter from Japan’s Communist Party’s central office. The DDR’s demand for the Chinese aggressors’ immediate

following section illustrates a more complex generic semantic grammar, built around the one formally presented here and implemented in the computer program Program for Linguistic Content Analysis, or PLCA (MetaText Inc. 1994). PLCA runs under DOS (version 2.2) and Windows 95 (version 3.0), and is available from iec PROGRAMMA (P.O. Box 841, 9700 AV GRONINGEN, the Netherlands, gamma.post@gamma.rug.nl).

21 In this English version the phrase, “said to be,” is used to translate a German tense (nonexistent in English) that designates the current sentence as a direct quotation from the most recently cited source. An appendix to this chapter gives the original German text of this news report.
unconditional withdrawal from Vietnam was affirmed yesterday in Berlin by Erich Honecker, General Secretary of the SED’s Central Committee and Chairman of the National Council: “The (East German) Republic shares the Socialist Republic of Vietnam’s government’s opinion that (indicates) vigilance is still called for, because Peking has tied its orders for withdrawal to new threats against Vietnam.” At an inspirational meeting last night in Helsinki, the Vietnamese Peace Committee’s chairman, Fan An, made clear his people’s confidence in victory over the Chinese invaders. “It grows from the fact,” he emphasized, “that all peace-loving humanity stands on our side.” (Berliner Rundfunk, 2nd news report, March 9, 1979)

CI coverage began in the sixth report of the January 29 morning news broadcast from the West Berlin radio station, Sender Freies Berlin (SFB), with a mention that Newsweek reported, “It is not to be ruled out that China plans a military strike against Vietnam.” No references to the event were made for at least one week after a minor aside on March 28 in which the East German radio station, Berliner Rundfunk (BR), mentioned “Peking-expansionists” in conjunction with official UN recognition of the Pol Pot Regime. In like fashion a continuous period of SR coverage (flanked by no SR coverage at least one week prior and following) occurred between May 21 and July 20. Two news broadcasts (one each from BR and SFB) were tape-recorded every morning during the period of study. Transcripts were made of all CI- or SR-related news reports within these broadcasts. Each clause within the reports was then identified by a fluent German speaker, yielding a total of 1923 clauses. A stratified random sample of

---

22 “Clause” is referred to here in its usual sense, namely as a sentence or part of a sentence that contains an inflected verb and, optionally, a subject or object, plus all modifiers related to this verb, subject, and object. Note that high coder reliability can be ensured only if coders initially agree on “what counts as a clause.” For example, coders must follow explicit rules regarding cases of ellipsis, as occurs with the omitted “indicates” parenthetically inserted into the quotation by Erich Honecker in the twelfth sentence. Here a coding rule might be established to ensure that constructions of the form, “the (or a) thought/idea/belief/opinion/etc. that [such-and-such]...”, should be encoded as relative clauses with the verb, to indicate, taking the such-and-such clause as its object (e.g., the belief, that indicates that [vigilance is called for]). A more general rule was that inflected verbs were not to be treated as clause-identifiers when they indicated the source of quoted information. Source and audience information were instead encoded as characteristics of each clause. Thus “reported” (by the ‘Voice of
400 clauses was drawn from this total, with 100 clauses being sampled from each station (BR or SFB) by event (CI or SR) combination.

Using key-word-in-context software, the texts of these clauses were scanned for frequently used words and phrases, which were then grouped into a more comprehensive vocabulary consisting of nominal and verbal coding categories. Although specific codes were assigned to some frequently mentioned entities (e.g., VIETNAM, SOMOSA, THE UNITED NATIONS, etc.), most coding categories corresponded to much broader classes (e.g., LATIN AMERICAN COUNTRY, and verbs such as to ENCOURAGE/FACILITATE and to VIOLENTLY ACT). Only very occasionally was vocabulary added once this vocabulary-development stage had been completed.

Two fluent German speakers encoded the sampled clauses directly from the German transcripts. During regular meetings, the coders met to jointly resolve differences in their encodings in the light of an evolving set of consistent coding rules. For example, one coding rule was, “‘Journalist’ is to be encoded as ‘genitive + noun = MEDIA’s PERSON’.” Because references to a medium commonly indicated its nation of origin, the rule implies that mention of a Chinese journalist would be encoded as genitive + genitive + noun = CHINA’s MEDIA’s PERSON. The software used in encoding the news data restricted users to a single genitive for each noun, however. 23 Because the encoding of national affiliation was

Vietnam’ in sentence 5) and “emphasized” (by Fan An in sentence 14) did not identify distinct clauses. Applying these coding rules, the above-quoted BR broadcast contains 19 clauses, each identified by its inflected verb (in italics).

23 This software was PLCA (see note 20 and below). In addition to allowing the encoding of only one genitive per noun, the program only allows the encoding of prepositional phrases when they identify the objects of intransitive verbs (e.g., CHINA VIOLENTLY ACTED toward a PLACE) or of verbs’ nominal forms (e.g., VIOLENCE toward a PLACE). PLCA also does not allow arbitrary encoding of adjectival or adverbial modifiers. In principle, there is no reason why in addition to subject-verb-object relations arbitrary sequences of modifiers might not also be encoded, even if only according to the surface grammatical syntax of the original text. Guided on the one hand by the impracticality of assigning unique codes to each word in one’s texts and on the other hand by a desire not to gloss over the texts’ salient properties, the number of coding categories is in practice usually restricted to a manageable amount. This restriction may involve dropping incidental modifiers from consideration or, when deemed necessary to capture salient properties of the text, incorporating them into the coding through the use of noun modifiers or more specific thematic categories. A case in point is the current illustration in which the need for a distinct coding category is circumvented by encoding journalist as the noun, PERSON, supplemented with the genitive, MEDIA (i.e., as MEDIA’s PERSON). In contrast, two coding categories were used to distinguish among “subjecting another to artillery fire,” “opening fire,” “returning to one’s barracks,” and “eating dinner,” despite the fact that all are actions. Within news
TABLE 1
Breakdown of Clauses by Station, Event, and Functional Form

<table>
<thead>
<tr>
<th>Functional form</th>
<th>Berliner Rundfunk (BR)</th>
<th>Sender Freies Berlin (SFB)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Chinese Invasion</td>
<td>Sandinista Revolution</td>
</tr>
<tr>
<td>Recognition</td>
<td>28</td>
<td>34</td>
</tr>
<tr>
<td>Perception</td>
<td>67</td>
<td>64</td>
</tr>
<tr>
<td>Evaluation</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Justification</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Sampled total</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Population total</td>
<td>403</td>
<td>853</td>
</tr>
</tbody>
</table>

deemed by the researchers to be of greater theoretical importance than personhood in a study of national alignments in the news, the coding rule was elaborated into the following: ‘‘Journalist’ is to be encoded as ‘genitive + noun = MEDIA’s PERSON’, except in cases where the nationality of the journalist’s medium (e.g., China) is known. In these cases, ‘journalist’ is to be encoded as CHINA’s MEDIA.”

The computer program, PLCA (see note 20), was used in applying vocabulary and coding rules to the sampled clauses. Although coders were blind to clauses’ dates and radio stations of origin, they were provided the entire news report within which each sampled clause appeared. Access to entire news reports provided coders with invaluable contextual information needed in determining sampled clauses’ intended meanings and corresponding functional forms. After each functional form was selected, PLCA guided coders through a series of prompts for vocabulary being assigned to the various syntactic components (e.g., subject, verb, etc.) within that form’s semantic grammar.

Table 1 provides a breakdown of the clauses by stations, events, and functional forms (i.e., recognition, perception, evaluation, and justification). One noteworthy finding in Table 1 is that of the sample’s eight judgmental clauses, seven were aired on East German radio. The probability is .035 that sampling error accounts for this finding, namely that positive or negative judgments are more likely to be conveyed in news broadcasts on BR than on SFB. Nonetheless, 98 percent of the sample’s clauses were reports of armed conflict, the violent nature of the former two clearly distinguished them from the latter two phrases. To capture this distinction, thematic categories were created for two classes of verbs, to VIOLENTLY ACT and to ACT.
descriptive of either processes or states of affairs—an understandable fact, given the descriptive nature of news reporting. The following subsection explains how the generic semantic grammar was applied to the sample’s 400 clauses. An illustrative log-linear model is then fit to data encoded from the 275 perception-clauses in the sample.

3.1. Encoding Clauses

The generic semantic grammar developed in the previous section is at the heart of Roberts’s (1989, 1991) quantitative text analysis method, Linguistic Content Analysis (LCA). Given the clause-specific relevance of the grammar, the method must be applied one clause at a time. Let us begin by encoding the first sentence in the BR transcript. In it is described the process of Chinese troops provoking fighting. During the encoding process, words such as “troops,” “provoke,” and “fighting” are likely to be encoded as falling into more general thematic categories. Classifying “troops” into the thematic category, “MILITARY/WEAPON,” “provoke” into the category, “ENCOURAGE/FACILITATE,” and “fighting” into the category, “VIOLENCE,” and applying the “(x)f(x)” functional form appropriate for this perception, the sentence might be rendered as “CHINA’s MILITARY/WEAPON ENCOURAGE/FACILITATES a VIOLENCE.”

The fourth sentence conveys a negative judgment of a Vietnamese train station’s and hospital’s (viz., two states of affairs’) inherent conditions. The syntax for a negatively stated evaluation such as this is “(x)a(x)Q_a,” and might be rendered as “VIETNAM’s PLACE is pitiful.” The fifth sentence is a direct quotation from the radio station, Voice of Vietnam. In it a withdrawal of troops (a state of affairs) is described as absent. If reference here were instead to a withdrawing of troops (a process), the sentence might best be encoded as a perception. Yet given the nominal, nonerund form of withdraw, it is more appropriately encoded as a recognition. Applying the corresponding functional form (see note 10), the sentence might be rendered as “VIETNAM’s MEDIA said, ‘There is not a WITHDRAWAL of a MILITARY/WEAPON.’ ” The eighth sentence describes the process of the Chinese doing harm (as in answer to the question, “What did they do?’”). Yet by applying the abstract judgmental label, “barbaric crimes,” to the harm-doing, the sentence also strongly conveys a negative judgment of this process (as in answer to the question, “Is it not terrible what they are doing?’”). Thus this coder encoded the sentence twice, once as a perception (rendered, “CHINA harmed.”) and once as a justification (rendered, “CHINA’s doing harm was ethically wrong.”). In acce-
dance with the generic semantic grammar, the syntax of this latter rendering is “(x)f(x)Q_p”.

The LCA encoding of judgmental clauses (i.e., evaluations and justifications) is more complex than the encoding of descriptive clauses (i.e., perceptions and recognitions). Research by Thomas and Heise (1995) suggests that culture and gender differences influence the affective meanings that individuals associate with words. Accordingly, coders should be particularly sensitive to the source’s world when the source is not of their culture or gender. Likewise, they must strive not to allow their own social biases to color their assignment of codes.

In addition to enhancing coder reflexivity, refinement of the generic semantic grammar’s qualifiers adds rigor to the process of coding judgmental clauses. Most text analysis methods provide two dimensions along which to encode affect: positive-negative and strong-weak. For example, the sentence, *The army was great*, appears (assuming lack of sarcasm) intended to convey a strong positive evaluation of the army. However, still unaccounted for is a third dimension, namely the type of evaluation being made. Without more contextual information, it is impossible to determine whether the inherent condition (e.g., health) or the utilitarian value (e.g., destructive capacity) of the army is being evaluated. Likewise, if *The army performed terribly* were intended to negatively judge the process of the army’s performing, it would remain unclear if this process were being condemned as inherently unethical, technically flawed, socially inept, or consequently ineffective. Like social historians who are obliged to address counterfactual queries, LCA coders must identify the types as well as the strength and sign of the evaluations and justifications in their texts.

No matter how structured the coding process, some coder disagreement is inevitable. For example, in the fifth sentence *a real withdrawal of the Peking troops* may be thought by some readers to represent a process rather than a state of affairs. As in the eighth sentence, a coder may decide that the clause was intended to have more than one meaning and that it should be encoded more than once. Disagreements among coders are, of course, the makings of poor interrater agreement. However, they are also “grist” in the development of explicit coding rules (for example, “Only noun phrases rendered as gerunds will be considered processes in selections of functional form.”). As it turns out, when coding rules have been

---

24 Multiple coding requires the assignment to each encoding of a weight equivalent to the proportion of its contribution to the clause’s overall meaning. To ensure sufficient interrater agreement, assignment will nearly always be of equivalent weights that sum to unity.
conscientiously developed and applied, coders are likely to attain consistently high agreement in LCA encodings (cf. Eltinge and Roberts 1993). One year after the encoding of this study’s 400 clauses, 20 of these were selected at random and encoded for a second time. The two encodings identified 11 perceptions and 7 recognitions in common, yielding significant evidence of agreement beyond chance ($\kappa = .79$, $p < .01$).\footnote{In addition, 19 of the 20 clauses (including all 11 clauses jointly identified as perceptions) had sources encoded identically both times. Beyond radio station, event, and source, the only additional variable included in the below statistical analysis (and, thus, for which reliability data need be provided) is the subject of the processes described in the data set’s 275 perception clauses. Two minor discrepancies occurred among the subject codes in the 11 clauses jointly identified as perceptions: COMMUNISM’s ORGANIZATION versus COMMUNISM’s TROOPS and THE ORGANIZATION OF AMERICAN STATES’ OFFICIAL versus THE UNITED STATES’ OFFICIAL. However, these discrepancies disappear with the collapsing of subject codes prior to analysis (i.e., when both of the former encodings are collapsed into an East-aligned category and both of the latter encodings are collapsed into a non-East-aligned category). Thus, insofar as the subsample of 11 perception clauses can be generalized to the 275 clauses analyzed below, rater agreement was effectively perfect once a clause’s functional form was determined.}

3.2. LCA Translation and PLCA ‘Retranslation’

Text analysts have frequently depicted their work as a process of translation (Andrén 1981; Franzosi 1994). Yet in making their translations, semantic text analysts will want to avoid both the literal rendering of surface grammar and the paraphrasing of text. For example, when used idiomatically, *I’ll stand by you in your misery* should not be encoded according to its surface grammar (i.e., as “I” and “you” having a relation of “standing together”). It would be equally inappropriate to paraphrase the idiom as “I assure you,” despite the fact that the sentence is an assurance. In contrast, an LCA translation renders the clause’s words according to the syntax of its appropriate functional form (i.e., not as a description of standing—a state of affairs, but as a description of assisting—a process). Thus, in a more accurate translation, the expression “to stand by” would be rendered as “to assist,” leaving my relation to you being one of promised future assistance. That is, an acceptable translation would be, “I shall assist you.”

The translation metaphor is not only apparent in linguistic content analysis, PLCA literally renders LCA encoding as a translation. The lefthand column in Table 2 provides a PLCA-generated translation of the entire transcript of the March 9 BR news report. By generating ongoing translation of the text being encoded, PLCA affords the coder ready verification that the themes in the original text are related correctly according
<table>
<thead>
<tr>
<th>LCA Rendering</th>
<th>English Translation from German</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHINA’s MILITARY/WEAPON ENCOURAGE/FACILITATES a VIOLENCE.</td>
<td>1 Despite their withdrawal-notifications on Monday, Peking’s aggression-troops still provoke new fighting.</td>
</tr>
<tr>
<td>CHINA’s MILITARY/WEAPON VIOLENTLY ACTED toward VIETNAM’s PLACE.</td>
<td>2 Thus the north Vietnamese provincial capital, Lan Shon, was subjected to the Chinese invaders’ sporadic artillery fire.</td>
</tr>
<tr>
<td>The MEDIA said, “CHINA’s MILITARY/WEAPON VIOLENTLY ACTED toward VIETNAM’s PLACE.</td>
<td>3 According to agency reports, shells struck the main streets flagrantly anew.</td>
</tr>
<tr>
<td>VIETNAM’s PLACE is pitiful.”</td>
<td>4 The train station and the provincial hospital are said to be completely destroyed.</td>
</tr>
<tr>
<td>VIETNAM’s MEDIA said, “There is not a WITHDRAWAL of a MILITARY/WEAPON.</td>
<td>5 As the station, “Voice of Vietnam,” reported, as yet there is still no indication of a real withdrawal of the Peking troops.</td>
</tr>
<tr>
<td>CHINA did somewhat WITHDRAW a MILITARY/WEAPON.</td>
<td>6 At a few locations the Chinese were said to have removed only certain ranks.</td>
</tr>
<tr>
<td>There was CHINA’s MILITARY/WEAPON.”</td>
<td>7 In other regions they were said still to hold their positions.</td>
</tr>
<tr>
<td>CHINA harmed.</td>
<td>8(a) During their movements the invaders continued to commit</td>
</tr>
<tr>
<td>CHINA’s doing harm was ethically wrong.</td>
<td>8(b) barbaric crimes.</td>
</tr>
<tr>
<td>VIETNAM’s MEDIA said, “CHINA does not DISCOURAGE/IMPEDE a VIOLENCE toward a MEDIA.”</td>
<td>9 According to statements from the news agency, VNA, China does not even resist attacking foreign journalists.</td>
</tr>
</tbody>
</table>
CHINA’s MILITARY/WEAPON VIOLENTLY ACTED toward a MEDIA’s PERSON(S), therefore COMMUNISM’s PERSON(S) was harmed.”

EAST GERMANY’s OFFICIAL DESCRIBED/REPORTED a DEMAND of CHINA.

EAST GERMANY’s OFFICIAL said, “EAST GERMANY’s BELIEF/FAITH is VIETNAM’s BELIEF/FAITH (that CONVEYS that [the PERCEPTION/CONSIDERATION is a necessity/certainty, because [according to what CHINA said] the obligation to WITHDRAW LOGICALLY RELATES to a THREAT to VIETNAM]).”

VIETNAM’s OFFICIAL DESCRIBED/REPORTED VIETNAM’s BELIEF/FAITH.

VIETNAM’s OFFICIAL said, “There becomes VIETNAM’s BELIEF/FAITH, because VIETNAM’s PERSON(S) is good for a good PERSON(S).”

10 Sharpshooters are said to have opened fire on a group of correspondents in Lan Shon, whereby they murdered Isamu Takan, a reporter from Japan’s Communist Party’s central office.

11 The DDR’s demand for the Chinese aggressors’ immediate unconditional withdrawal from Vietnam was affirmed yesterday in Berlin by Erich Honnecker, General Secretary of the SED’s Central Committee and Chairman of the National Council.

12 “The (East German) Republic shares the Socialist Republic of Vietnam’s government’s opinion that (indicates) vigilance is still called for, because Peking has tied its orders for withdrawal to new threats against Vietnam.”

13 At an inspirational meeting last night in Helsinki, the Vietnamese Peace Committee’s chairman, Fan An, made clear his people’s confidence in victory over the Chinese invaders.

14 “It grows from the fact,” he emphasized, “that all peace-loving humanity stands on our side.”

---

Note: This news report was broadcast by the East Berlin radio station, Berliner Rundfunk, on the morning of March 9, 1979. The LCA rendering in this table was generated by PLCA (MetaText, Inc.). The original German version of the news report is in an appendix to this chapter. In the English translation, the phrase, “credited to,” is used in translating a German tense (nonexistent in English) that designates the current sentence as a direct quotation from the most recently cited source.
to the functional forms appropriate to their respective clauses. Yet the purpose of such semantic encoding is not merely to translate text into a sort of unambiguous pidgin English (a.k.a. LCAese). Like all quantitative text analysis methods, LCA’s encoding process is intended to produce a data matrix suitable for statistical analysis.

The translation in Table 2 is in fact a retranslation. That is, it is a direct translation of a data matrix, and only indirectly a translation of the original text. Transparent to the user, the encoding process involves a translation from text to numbers (i.e., to an unambiguous matrix representation). The user is provided the retranslation to make the same kind of verification that developers of machine translation software use to check their routines: After translating the same text from one language to another, the translation is retranslated back to the original. If the retranslation closely resembles the original, the software’s functionality is verified.

Unlike translation software, PLCA automates the retranslation but not the translation process. In translating verbatim text into a data matrix, the program’s role is one of guiding the user through a sequence of coding decisions. Advances in the fields of linguistics and artificial intelligence remain insufficient to automate the coder’s ability to recognize texts’ intended meanings (cf. Shapiro 1997). Thus, although potentially supported by various software amenities, the translation from words to numbers requires that coders make considerable use of both their intuition of the source’s meanings and their knowledge of the situational context.

The purpose of PLCA’s retranslation capability (i.e., from data matrix to LCAese) is to enable coders to verify whether their intuitions are accurately represented in a semantically unambiguous linguistic form that is grounded in the generic semantic grammar described above. LCA encoding requires more of coders than that they merely identify subjects and predicates; they must commit themselves to unambiguous encodings of oftentimes ambiguous linguistic expressions. By comparing original text to LCAese, one is able to evaluate whether the retranslation captures one’s intuition that a clause was intended to communicate a description or judgment of a process or state of affairs.

Although PLCA renders linguistic data into a language comprehensible to English speakers, this language is not English but LCAese—a language that unlike English, allows each of a speech act’s intended meanings to be expressed in one and only one way. That is, unlike natural languages, LCAese is a language without ambiguity. As a consequence, users cannot legitimately disagree about the meaning of an LCAese expression, only about how well distinct LCAese expressions render the intended meaning(s) of a particular speech act.
Table 3 lists the data matrix from which the retranslation in Table 2 was generated. The first five columns of the matrix give identifying and sequencing information. Missing data are rendered as periods. The first clause (encoded in the first row of the matrix) is a perception (functional form = 1) in present tense (clause tense = 2), that is not a question (question = 0) and that functions to convey a process whereby CHINA’s (genitive of subject = 213) MILITARY/WEAPON (subject = 239) ENCOURAGE/FACILITATES (clause valence = T, main verb = 33) VIOLENCE (object = 50). These codes along with the LCA syntax for a perception comprise sufficient information for LCA software to render the first clause as it appears in Table 2. Other retractions can be reconstructed via similar comparisons between the two tables.\textsuperscript{27}

\textsuperscript{27}Interpretation of the data in Table 3 will be easier if more code definitions are provided: Identifying information, clause number, and sentence number should be self-explanatory. A main clause has no (i.e., zero) depth in its sentence’s syntax tree; a clause subordinate to the main clause has a depth of one; a clause subordinate to this subordinate clause has a depth of two; and so on. There are three forms of subordination. A conjunctive clause is subordinated to a main clause via a conjunction (e.g., T = therefore or 6 = because). The subordination code, OR, identifies a clause that is relative to the object of the previous clause (e.g., “There is a belief that indicates a need for vigilance”). The subordination code, OP, identifies an object proxy clause that is itself the object of the previous clause (e.g., “The belief indicates that vigilance is necessary”). Valence codes (i.e., indicators of semantic opposition) differ among functional forms. Among perceptions and recognitions, the code, T, identifies a clause that is positively stated. Negation of the main verb in perception and recognition clauses can be identified with NF and SI codes, respectively calling for retractions in LCAese that include the words not (e.g., “There is not a WITHDRAWAL”) and somewhat (e.g., “CHINA did somewhat WITHDRAW”). In an evaluation of a state of affairs, the valence codes, - and +, are rendered in LCAese respectively with the predicate adjectives pitiful (e.g., “The PLACE is pitiful”) and good (e.g., “VIETNAM is good”). In a justification of a process, the valence code, E-, is rendered in LCAese using ethically wrong (e.g., “CHINA’s doing harm was ethically wrong”). Each nominal syntactic component (i.e., audience, source, subject, and object) has a genitive associated with it. When both noun and genitive are phenomena (i.e., having codes > 101), genitives in LCAese take possessive forms such as “CHINA’s MILITARY” (genitive = 213, noun = 239) or “VIETNAM’s PLACE” (genitive = 214, noun = 120). When the genitive is the nominal form of a verb, in LCAese the noun becomes its object (e.g., “DEMAND of CHINA” [genitive = 28, noun = 213] and “OBLIGATION to WITHDRAW” [genitive = 3, noun = 56]). Whereas clauses 5, 7, 14, and 16 are recognition clauses with main verb codes of 1 (i.e., “is an instance of”), clause 19 is a recognition clause with a main verb code of 2 (i.e., “becomes”). (See footnote 10 for more detail on the four types of recognition clauses.) The three last columns indicate whether clause-weights are fractions of sentence-weights, what the base-weight is for all clauses with the same identifying information, and what the clause-weight is after taking a clause-specific weight and the other weight information into account.
<table>
<thead>
<tr>
<th>Identifying information</th>
<th>Clause number</th>
<th>Sentence number</th>
<th>Depth in syntax tree (0=main clause, 1=subord. to main, etc.)</th>
<th>Functional form (P=1, R=2, J=3, or E=4)</th>
<th>Type of subordination (conjunctive, relative, or proxy)</th>
<th>Clause tense (1=past, 2=present, 3=future)</th>
<th>Question? (1=where, 2=when, etc.)</th>
<th>Clause valence (positive or negative, good or bad)</th>
<th>Genitive of audience</th>
<th>Audience</th>
<th>Genitive of source</th>
<th>Source</th>
<th>Genitive of subject</th>
<th>Subject</th>
<th>Modal auxiliary verb</th>
<th>Main verb</th>
<th>Genitive of object</th>
<th>Object</th>
<th>Weight info.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ECHINA</td>
<td>MAR09</td>
<td>21</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>T</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td></td>
<td>.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>.</td>
<td>213</td>
<td>239</td>
<td>.</td>
<td>50</td>
<td>C</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ECHINA</td>
<td>MAR09</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>T</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>213</td>
<td>239</td>
<td>.</td>
<td>50</td>
<td>214</td>
<td>120</td>
<td>C</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ECHINA</td>
<td>MAR09</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>.</td>
<td>1</td>
<td>0</td>
<td>T</td>
<td>.</td>
<td>.</td>
<td>240</td>
<td>213</td>
<td>239</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>50</td>
<td>214</td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>---</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>---</td>
<td>---</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>----</td>
</tr>
<tr>
<td>ECHINA</td>
<td>MAR09</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>.</td>
<td>2</td>
<td>0</td>
<td>-</td>
<td>.</td>
<td>.</td>
<td>240</td>
<td>214</td>
<td>120</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>C</td>
</tr>
<tr>
<td>ECHINA</td>
<td>MAR09</td>
<td>5</td>
<td>5</td>
<td>0</td>
<td>2</td>
<td>.</td>
<td>2</td>
<td>0</td>
<td>NI</td>
<td>.</td>
<td>214</td>
<td>240</td>
<td>.</td>
<td>.</td>
<td>1</td>
<td>56</td>
<td>239</td>
<td>C</td>
<td>1.00</td>
</tr>
<tr>
<td>ECHINA</td>
<td>MAR09</td>
<td>6</td>
<td>6</td>
<td>0</td>
<td>1</td>
<td>.</td>
<td>1</td>
<td>0</td>
<td>SI</td>
<td>.</td>
<td>214</td>
<td>240</td>
<td>.</td>
<td>213</td>
<td>.</td>
<td>56</td>
<td>.</td>
<td>239</td>
<td>C</td>
</tr>
<tr>
<td>ECHINA</td>
<td>MAR09</td>
<td>7</td>
<td>7</td>
<td>0</td>
<td>2</td>
<td>.</td>
<td>1</td>
<td>0</td>
<td>T</td>
<td>.</td>
<td>214</td>
<td>240</td>
<td>.</td>
<td>.</td>
<td>1</td>
<td>213</td>
<td>239</td>
<td>C</td>
<td>1.00</td>
</tr>
<tr>
<td>ECHINA</td>
<td>MAR09</td>
<td>8</td>
<td>8</td>
<td>0</td>
<td>1</td>
<td>.</td>
<td>1</td>
<td>0</td>
<td>T</td>
<td>.</td>
<td>.</td>
<td>213</td>
<td>.</td>
<td>99</td>
<td>.</td>
<td>.</td>
<td>C</td>
<td>1.00</td>
<td>0.50</td>
</tr>
<tr>
<td>ECHINA</td>
<td>MAR09</td>
<td>9</td>
<td>9</td>
<td>0</td>
<td>3</td>
<td>.</td>
<td>1</td>
<td>0</td>
<td>E®</td>
<td>.</td>
<td>.</td>
<td>213</td>
<td>.</td>
<td>99</td>
<td>.</td>
<td>.</td>
<td>C</td>
<td>1.00</td>
<td>0.50</td>
</tr>
<tr>
<td>ECHINA</td>
<td>MAR09</td>
<td>10</td>
<td>10</td>
<td>0</td>
<td>1</td>
<td>.</td>
<td>2</td>
<td>0</td>
<td>NI</td>
<td>.</td>
<td>214</td>
<td>240</td>
<td>.</td>
<td>213</td>
<td>.</td>
<td>34</td>
<td>50</td>
<td>240</td>
<td>C</td>
</tr>
<tr>
<td>ECHINA</td>
<td>MAR09</td>
<td>11</td>
<td>11</td>
<td>0</td>
<td>1</td>
<td>.</td>
<td>1</td>
<td>0</td>
<td>T</td>
<td>.</td>
<td>214</td>
<td>240</td>
<td>.</td>
<td>213</td>
<td>.</td>
<td>50</td>
<td>240</td>
<td>200</td>
<td>C</td>
</tr>
<tr>
<td>ECHINA</td>
<td>MAR09</td>
<td>12</td>
<td>12</td>
<td>1</td>
<td>1</td>
<td>.</td>
<td>1</td>
<td>0</td>
<td>T</td>
<td>.</td>
<td>214</td>
<td>240</td>
<td>.</td>
<td>.</td>
<td>99</td>
<td>116</td>
<td>200</td>
<td>C</td>
<td>1.00</td>
</tr>
<tr>
<td>ECHINA</td>
<td>MAR09</td>
<td>13</td>
<td>13</td>
<td>0</td>
<td>1</td>
<td>.</td>
<td>1</td>
<td>0</td>
<td>T</td>
<td>.</td>
<td>.</td>
<td>211</td>
<td>201</td>
<td>.</td>
<td>24</td>
<td>28</td>
<td>213</td>
<td>C</td>
<td>1.00</td>
</tr>
<tr>
<td>ECHINA</td>
<td>MAR09</td>
<td>14</td>
<td>14</td>
<td>0</td>
<td>2</td>
<td>.</td>
<td>2</td>
<td>0</td>
<td>T</td>
<td>.</td>
<td>211</td>
<td>201</td>
<td>.</td>
<td>211</td>
<td>17</td>
<td>.</td>
<td>1</td>
<td>214</td>
<td>17</td>
</tr>
<tr>
<td>ECHINA</td>
<td>MAR09</td>
<td>15</td>
<td>15</td>
<td>1</td>
<td>1</td>
<td>.</td>
<td>2</td>
<td>0</td>
<td>T</td>
<td>.</td>
<td>211</td>
<td>201</td>
<td>214</td>
<td>17</td>
<td>20</td>
<td>.</td>
<td>.</td>
<td>C</td>
<td>1.00</td>
</tr>
<tr>
<td>ECHINA</td>
<td>MAR09</td>
<td>16</td>
<td>16</td>
<td>2</td>
<td>2</td>
<td>.</td>
<td>2</td>
<td>0</td>
<td>T</td>
<td>.</td>
<td>211</td>
<td>201</td>
<td>.</td>
<td>15</td>
<td>1</td>
<td>.</td>
<td>.</td>
<td>C</td>
<td>1.00</td>
</tr>
<tr>
<td>ECHINA</td>
<td>MAR09</td>
<td>17</td>
<td>17</td>
<td>3</td>
<td>1</td>
<td>.</td>
<td>6</td>
<td>2</td>
<td>0</td>
<td>T</td>
<td>.</td>
<td>.</td>
<td>213</td>
<td>3</td>
<td>56</td>
<td>.</td>
<td>41</td>
<td>29</td>
<td>214</td>
</tr>
<tr>
<td>ECHINA</td>
<td>MAR09</td>
<td>18</td>
<td>18</td>
<td>4</td>
<td>0</td>
<td>.</td>
<td>1</td>
<td>0</td>
<td>T</td>
<td>.</td>
<td>.</td>
<td>214</td>
<td>201</td>
<td>.</td>
<td>24</td>
<td>214</td>
<td>17</td>
<td>C</td>
<td>1.00</td>
</tr>
<tr>
<td>ECHINA</td>
<td>MAR09</td>
<td>19</td>
<td>19</td>
<td>5</td>
<td>0</td>
<td>.</td>
<td>2</td>
<td>0</td>
<td>T</td>
<td>.</td>
<td>214</td>
<td>201</td>
<td>.</td>
<td>.</td>
<td>2</td>
<td>214</td>
<td>17</td>
<td>C</td>
<td>1.00</td>
</tr>
<tr>
<td>ECHINA</td>
<td>MAR09</td>
<td>20</td>
<td>20</td>
<td>1</td>
<td>4</td>
<td>6</td>
<td>2</td>
<td>0</td>
<td>+</td>
<td>.</td>
<td>214</td>
<td>201</td>
<td>214</td>
<td>200</td>
<td>.</td>
<td>98</td>
<td>200</td>
<td>C</td>
<td>1.00</td>
</tr>
</tbody>
</table>

*Source:* The output in this table was generated by PLCA (MetaText, Inc.).
3.3. A Statistical Analysis of the Encoded Data

In analyzing LCA data one must be careful not to simultaneously analyze clauses encoded according to distinct functional forms. To do so would be to mix proverbial apples and oranges, leading to not-so-mixed-metaphorical garbage-in, garbage-out consequences. (More on this at the close of the next subsection.) Accordingly, the exploratory analysis in this subsection is restricted to the 275 perception-clauses in the data set. That is, the analysis is of processes described in East and West Berlin radio news coverage of two military conflicts.

To take the structure of the data set into account, terms for radio station (i.e., BR or SFB), event (i.e., CI or SR), and their interaction must be included in the analysis. In addition, let us consider which quoted source (i.e., a person, organization, or anthropomorphized country that is East-aligned, non–East-aligned, 28 or neutral, or the case of no quoted source) describes which subject (i.e., a person, organization, or country that is East-aligned, non–East-aligned, or neutral, or a nonvolitional thing such as a place, or the case of no subject given, as in passive voice) as the initiator of a process. Table 4 lists a series of hierarchical log-linear models in which various joint distributions among these four variables are fit to the LCA data.

The baseline model, Model (4), provides the best fit, given that its BIC statistic is the smallest among Models (1) through (3), in which a single term was added to the baseline model, and Models (5) through (7) in which a single term was dropped. 29 That is, no improvement in fit is gained by adding to this equation terms for QS, SE, or RQE interactions, and that a loss of fit results by removing RS, QE, or RQ from the equation. Model (4) has the following linear form. 30

28 Under non–East-aligned fall the non–East-aligned Chinese and the West-aligned Somosan parties to the respective conflicts.

29 Raftery (1986a, 1986b) has proposed the Bayesian information criterion (or BIC) as an alternative to likelihood ratio tests for selecting among hierarchical log-linear models. The smaller the value of BIC, the better the fit. Although the [RE] interaction affords no improvement in fit, it is retained in all models to ensure that parameter estimates are conditioned for structure in the data resulting from the stratified sampling design.

30 When a clause was broadcast by BR, i = 1, when broadcast by SFB, i = 2; when the source of a clause is East-aligned, j = 1, when non–East-aligned, j = 2, when neutral, j = 3, when no source is given, j = 4; when the subject of a clause is East-aligned, k = 1, when non–East-aligned, k = 2, when neutral, k = 3, when a thing or place, k = 4, when no subject is given, k = 5; when a clause is about the Chinese invasion, l = 1, when about the Sandinista Revolution, l = 2.
TABLE 4
Likelihood-Ratio Chi-Squares ($L^2$) for Hierarchical Loglinear Models

<table>
<thead>
<tr>
<th>Model</th>
<th>Marginals Fitted</th>
<th>$df$</th>
<th>$L^2$</th>
<th>$p$</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>[RQ][RS][RE][QS][QE]</td>
<td>47</td>
<td>52.98</td>
<td>.254</td>
<td>$-211.01$</td>
</tr>
<tr>
<td>(2)</td>
<td>[RQ][RS][RE][OE][SE]</td>
<td>55</td>
<td>64.22</td>
<td>.185</td>
<td>$-244.70$</td>
</tr>
<tr>
<td>(3)</td>
<td>[RQE][RS]</td>
<td>57</td>
<td>68.30</td>
<td>.145</td>
<td>$-251.86$</td>
</tr>
<tr>
<td>(4)</td>
<td>[RQ][RS][RE][OE] (baseline)</td>
<td>59</td>
<td>77.21</td>
<td>.056</td>
<td>$-254.18$</td>
</tr>
<tr>
<td>(5)</td>
<td>[RQ][S][RE][OE]</td>
<td>63</td>
<td>101.45</td>
<td>.002</td>
<td>$-252.41$</td>
</tr>
<tr>
<td>(6)</td>
<td>[RQ][S][RE]</td>
<td>62</td>
<td>101.26</td>
<td>.001</td>
<td>$-246.98$</td>
</tr>
<tr>
<td>(7)</td>
<td>[RS][RE][OE]</td>
<td>62</td>
<td>111.71</td>
<td>&lt;.001</td>
<td>$-236.53$</td>
</tr>
</tbody>
</table>

Note. R = Radio station, Q = Quoted source, S = Semantic subject, E = Event. The [RE] interaction is included in all models to adjust for the stratified sampling design.

\[
\ln(\hat{F}_{ijkl}) = \eta + \lambda_i^R + \lambda_j^Q + \lambda_k^S + \lambda_l^E + \lambda_{ij}^{RQ} + \lambda_{ik}^{RS} + \lambda_{il}^{RE} + \lambda_{jl}^{QE}
\]

where

- $i$ represents the two categories of radio station (R),
- $j$ represents the four categories of quoted source (Q),
- $k$ represents the five categories of semantic subject (S), and
- $l$ represents the two categories of event (E).

Table 5 lists the parameter estimates for Model (4). Significantly nonzero estimates of interactions with radio station yield findings on the rhetoric of political allegiance. For example, the interaction parameter associated with joint occurrences of Berliner Rundfunk radio station and the quotation of an East-aligned source is $\lambda_{ij(11)}^{RQ} = .59$. Conditional on other effects in the model, the log odds are four times this (i.e., 2.36) that a clause broadcast on Berliner Rundfunk mentioned an East-aligned (rather than another or no) person, organization, or country as the source being quoted, or that a clause on Sender Freies Berlin did not (rather than did) mention an East-aligned source. Further restricting this discussion to significant interaction parameters, in Table 5 one also finds the odds to be significantly greater that SFB (and less that BR) mentioned non-East-aligned persons, organizations, and countries either as the source of a quotation or as the initiator of a process. Thus it was in these ways that East and West Berlin radio stations cited and reported on parties with which their nations were politically aligned.

News coverage of the Chinese invasion of Vietnam contained a disproportionately large number of quotations from Eastern sources. That
### TABLE 5
Parameter Estimates and Standard Errors for Model 4

<table>
<thead>
<tr>
<th>Description</th>
<th>Estimate</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Marginal Effects</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Station effects</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a. Berliner Rundfunk</td>
<td>.027</td>
<td>.124</td>
</tr>
<tr>
<td>b. Sender Freies Berlin</td>
<td>-.027</td>
<td>.124</td>
</tr>
<tr>
<td>2. Source effects</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a. East-aligned</td>
<td>-.367</td>
<td>.189</td>
</tr>
<tr>
<td>b. Non-East-aligned</td>
<td>-.526</td>
<td>.214</td>
</tr>
<tr>
<td>c. Neutral</td>
<td>-.573</td>
<td>.184</td>
</tr>
<tr>
<td>d. None given</td>
<td>1.466</td>
<td>.117</td>
</tr>
<tr>
<td>3. Subject effects</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a. East-aligned</td>
<td>.452</td>
<td>.141</td>
</tr>
<tr>
<td>b. Non-East-aligned</td>
<td>.919</td>
<td>.127</td>
</tr>
<tr>
<td>c. Neutral</td>
<td>.391</td>
<td>.141</td>
</tr>
<tr>
<td>d. Other (e.g., a place or thing)</td>
<td>-1.732</td>
<td>.331</td>
</tr>
<tr>
<td>e. None given (i.e., passive voice)</td>
<td>-.031</td>
<td>.159</td>
</tr>
<tr>
<td>4. Event effects</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a. Chinese invasion</td>
<td>.093</td>
<td>.087</td>
</tr>
<tr>
<td>b. Nicaragua revolution</td>
<td>-.093</td>
<td>.087</td>
</tr>
<tr>
<td><strong>Interactions with Station</strong>a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Station by quoted source effects</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a. Berliner Rundfunk and East-aligned source</td>
<td>.590</td>
<td>.163</td>
</tr>
<tr>
<td>b. Berliner Rundfunk and non-East-aligned source</td>
<td>-.950</td>
<td>.211</td>
</tr>
<tr>
<td>c. Berliner Rundfunk and neutral source</td>
<td>.257</td>
<td>.179</td>
</tr>
<tr>
<td>d. Berliner Rundfunk and no source given</td>
<td>.102</td>
<td>.112</td>
</tr>
<tr>
<td>6. Station by semantic subject effects</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a. Berliner Rundfunk and East-aligned subject</td>
<td>.234</td>
<td>.141</td>
</tr>
<tr>
<td>b. Berliner Rundfunk and non-East-aligned subject</td>
<td>-.498</td>
<td>.127</td>
</tr>
<tr>
<td>c. Berliner Rundfunk and neutral subject</td>
<td>-.032</td>
<td>.141</td>
</tr>
<tr>
<td>d. Berliner Rundfunk and other subject</td>
<td>.449</td>
<td>.331</td>
</tr>
<tr>
<td>e. Berliner Rundfunk and no subject given</td>
<td>-.153</td>
<td>.159</td>
</tr>
<tr>
<td><strong>Interactions with Event</strong>b</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Event by station effects</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a. Chinese invasion and Berliner Rundfunk</td>
<td>.023</td>
<td>.067</td>
</tr>
<tr>
<td>b. Chinese invasion and Sender Freies Berlin</td>
<td>-.023</td>
<td>.067</td>
</tr>
<tr>
<td>8. Event by quoted source effects</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a. Chinese invasion and East-aligned source</td>
<td>.672</td>
<td>.174</td>
</tr>
<tr>
<td>b. Chinese invasion and non-East-aligned source</td>
<td>-.075</td>
<td>.152</td>
</tr>
<tr>
<td>c. Chinese invasion and neutral source</td>
<td>-.318</td>
<td>.175</td>
</tr>
<tr>
<td>d. Chinese invasion and no source given</td>
<td>-.279</td>
<td>.102</td>
</tr>
</tbody>
</table>

aParameter estimates for interactions with Sender Freies Berlin are the negatives of those with Berliner Rundfunk.

bParameter estimates for interactions with the Sandinista Revolution are the negatives of those with the Chinese invasion of Vietnam.
is, parameter estimates associated with the source-by-event interaction, 
[QE], show that significantly more clauses on CI than those on SR cited an 
Eastern source. In place of these Eastern citations, coverage of the San-
dinista Revolution (in comparison to CI coverage) simply had more clauses 
unattributed to any source. Because the Chinese invasion was a commu-
nist conflict (i.e., a conflict between communist nations with utopian ob-
jectives that are presumably the same), both BR and SFB journalists may 
have believed Eastern sources were needed to explain this “Eastern event” 
to their respective audiences.

These are the sorts of inferences afforded when randomly sampled 
texts are encoded according to a generic semantic grammar. Each of the 
previous two paragraphs contains assertions (supported at the .05 signi-
ficance level) about the relative frequencies of various types of clauses 
among the population of all clauses broadcast in two radio stations’ cov-
erage of two military conflicts. Moreover, these types of clauses are de-
defined, in part, according to specific relations among the semantic 
components of the clauses themselves.

3.4. Inferences from Thematic Versus Semantic Text Analysis

The inferences drawn in the previous subsection are different from those 
afforded in more traditional thematic text analyses (e.g., Namenwirth 1973; 
Stone 1997; Stone et al. 1966; Weber 1990). For comparative purposes, the 
Berlin radio news data are now reanalyzed using traditional coding meth-
ods, retaining the clause as unit of analysis, and analyzing the same 275 
perception clauses as before. Table 6 presents the data analyzed in the 
previous subsection; Table 7 provides data on the same clauses but en-
coded thematically. Both tables indicate how many clauses are associated 
with each combination of station and event. However, the semantically 
encoded data specify what the source and subject are for each clause, 
whereas the thematically encoded data list what combination of East, non-
East, and neutral affiliates are mentioned (as audience, source, subject, or 
object) in each clause. For example, among clauses broadcast on Berliner 
Rundfunk about the Chinese Invasion, four clauses mentioned no person, 
organization, or country with any of the three (East, non-East, or neutral) 
types of alignment; 17 clauses mentioned only East-aligned person(s), 
organization(s), or country(ies); 11 mentioned both East-aligned and non–
<table>
<thead>
<tr>
<th>Semantic Subject</th>
<th>Source Quoted on Berliner Rundfunk</th>
<th>Source Quoted on Sender Freies Berlin</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>East</td>
<td>Non-East</td>
</tr>
<tr>
<td>East-aligned</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>Non-East-aligned</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>Neutral</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Other</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>None</td>
<td>6</td>
<td>0</td>
</tr>
</tbody>
</table>

**Chinese Invasion of North Vietnam**

**Sandinista Revolution**
TABLE 7
Cross-Classification of Clauses According to Event, Station, and Occurrences of Various Combinations of Eastern, non-Eastern, and Neutral Alignment

<table>
<thead>
<tr>
<th>Pattern of Occurrences</th>
<th>Chinese Invasion</th>
<th>Sandinista Revolution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Berliner Rundfunk</td>
<td>Sender Freies Berlin</td>
</tr>
<tr>
<td></td>
<td>Berliner Rundfunk</td>
<td>Sender Freies Berlin</td>
</tr>
<tr>
<td>East</td>
<td>non-East</td>
<td>neutral</td>
</tr>
<tr>
<td>V</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>V</td>
<td>17</td>
<td>6</td>
</tr>
<tr>
<td>X</td>
<td>4</td>
<td>20</td>
</tr>
<tr>
<td>X</td>
<td>14</td>
<td>10</td>
</tr>
<tr>
<td>X</td>
<td>11</td>
<td>18</td>
</tr>
<tr>
<td>X</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>X</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>X</td>
<td>5</td>
<td>0</td>
</tr>
</tbody>
</table>

East-aligned but not neutral person(s), organization(s), or country(ies); and so on.

Using the same modeling procedure as in the previous section, the following log-linear model was determined to best fit the data in Table 7:

\[
\ln(\hat{\Phi}_{ijklm}) = \eta + \lambda_i^R + \lambda_j^A + \lambda_k^O + \lambda_l^U + \lambda_m^E + \lambda_{ij}^{RA} + \lambda_{ik}^{RO} + \lambda_{im}^{RE} + \lambda_{jk}^{AO} + \lambda_{jl}^{AU} + \lambda_{kl}^{OU} + \lambda_{lm}^{UE}
\]

where

- \( i \) represents the two categories of radio station (R),
- \( j \) represents the two categories of Eastern alignment (A),
- \( k \) represents the two categories of non-Eastern alignment (O),
- \( l \) represents the two categories of neutral alignment (U), and
- \( m \) represents the two categories of event (E).

\[31\] Tables for this analysis (i.e., ones similar to Tables 4 and 5) are available from the author on request.

\[32\] When a clause was broadcast by BR, \( i = 1 \), when broadcast by SFB, \( i = 2 \); when a clause mentions Eastern alignment, \( j = 1 \), otherwise \( j = 2 \); when a clause mentions non-Eastern alignment, \( k = 1 \), otherwise \( k = 2 \); when a clause mentions neutral alignment, \( l = 1 \), otherwise \( l = 2 \); when a clause is about the Chinese invasion, \( m = 1 \), when about the Sandinista Revolution, \( m = 2 \).
A lengthy discussion of this model’s parameters is not necessary, given that my purpose in this subsection is solely to distinguish the types of inferences afforded in a thematic versus a semantic text analysis. In fact, a comparison of one parameter from each model will suffice for this purpose.

The $\lambda_{ij(11)}$ parameter from the previous analysis has affinities with the $\lambda_{ij(11)}^{RA}$ parameter in this one. The former statistic allowed the inference that the log odds were 2.36 “that a clause broadcast on BR mentioned an East-aligned (rather than another, or no) person, organization, or country as the source being quoted.” Given that $\lambda_{ij(11)}^{RO} = .23$, the thematic text analysis affords the inference that the log odds were .92 “that a clause broadcast on BR mentioned an East-aligned (rather than another, or no) person, organization, or country.” The two inferences are substantively identical with the exception of the five words italicized at the end of the former. In brief, thematic text analysis affords inferences about the occurrence of various themes (or concepts) in texts; semantic text analysis affords inferences about how concepts are used within the framework of a (generic or other) semantic grammar.

Finally, it is important to note that in order to make this thematic text analysis comparable to the previous, semantic one, only perception clauses were analyzed. More commonly, all sampled text blocks (here all 400 sampled clauses) would have been scanned for theme occurrences. Changing emphasis for a moment, imagine that all 400 semantically encoded clauses are subjected to a statistical analysis. Superficially such an analysis has appeal because it would allow the addition of functional form to the variables in one’s model. Now further imagine that the analysis reveals a relatively large number of BR clauses having an East-aligned subject. The only legitimate interpretation of this finding would be that the East German station (BR) mentioned East-aligned persons, organizations, or countries more often than the West German station (SFB) either as initiators of processes (among perception clauses), as attributes of some object (among recognition clauses), as initiators of positively or negatively judged processes (among justification clauses), or as itself the subject of a positive or negative judgment (among evaluation clauses). Clearly, such an interpretation would afford little more than could have been ascertained from a thematic text analysis. (In fact, less would be gained because only subject segments would have been scanned.) This is why statistical models of LCA data must generally be estimated using data encoded according to a single functional form.
4. CONCLUSION

This paper describes and then applies a text-encoding method for researchers who wish to draw probabilistic inferences about the prominence of semantically related themes within the clauses of a population of texts. Linguists have correctly noted that syntax grammars map themes’ surface-grammatical relations and thus afford only ambiguous indicators of texts’ intended meanings. In contrast, the generic semantic grammar described here can be used to generate an unambiguous encoding of any clause with identifiable meaning(s). Once the coder has divined a clause’s meaning, the clause’s themes and theme relations are fit into the functional form appropriate to this meaning. The semantic grammar consists of four such unambiguous functional forms, each of which has a unique semantic opposite. In addition, the semantic grammar affords a sufficiently fine-grained mapping of text that the validity of the encoded data can be evaluated by retranslating them back into a semblance of the original.

The central message of a recent collection on quantitative text analysis methods is that researchers must select the text analysis method that most closely matches the research question at hand (Roberts 1997). If one’s objective is to use texts as a source of information about relational characteristics of some well-defined phenomenon (the event, the grievance, etc.), a phenomenal approach to semantic text analysis may be most appropriate (Franzosi 1994; Shapiro and Markoff, forthcoming). If one is primarily concerned with theme occurrences (but not theme relations), thematic text analysis may provide the best match (Stone 1997; Weber 1990). If one’s research question requires that text blocks be depicted as sets of interrelated themes, a network text analysis method may be what is needed (Carley 1993; Kleinnijenhuis, de Ridder, and Rietberg 1997). The method presented here is for addressing research questions on how themes are related in relatively unstructured text populations—questions often of interest to students of culture and the media.

APPENDIX

Die Aggressionstruppen Pekings provozieren ungeachtet ihrer Rückzugs- einkündigungen vom Montag, immer neue Kämpfe. So lag die nordviet-

REFERENCES


