Automatic Detection of Text Genre

Brett Kessler **Geoffrey Nunberg** **Hinrich Schütze**

Xerox Palo Alto Research Center
3333 Coyote Hill Road
Palo Alto CA 94304 USA

Department of Linguistics
Stanford University
Stanford CA 94305-2150 USA

email: {bkessler,nunberg,schuetze}@parc.xerox.com
URL: ftp://parcftp.xerox.com/pub/qca/papers/genre

Abstract

As the text databases available to users become larger and more heterogeneous, genre becomes increasingly important for computational linguistics as a complement to topical and structural principles of classification. We propose a theory of genres as bundles of *facets*, which correlate with various surface cues, and argue that genre detection based on surface cues is as successful as detection based on deeper structural properties.

1 Introduction

Computational linguists have been concerned for the most part with two aspects of texts: their *structure* and their *content*. That is, we consider texts on the one hand as formal objects, and on the other as symbols with semantic or referential values. In this paper we want to consider texts from the point of view of genre; that is, according to the various functional roles they play.

Genre is necessarily a heterogeneous classificatory principle, which is based among other things on the way a text was created, the way it is distributed, the register of language it uses, and the kind of audience it is addressed to. For all its complexity, this attribute can be extremely important for many of the core problems that computational linguists are concerned with. Parsing accuracy could be increased by taking genre into account (for example, certain object-less constructions occur only in recipes in English). Similarly for POS-tagging (the frequency of uses of *trend* as a verb in the *Journal of Commerce* is 35 times higher than in *Sociological Abstracts*). In word-sense disambiguation, many senses are largely restricted to texts of a particular style, such as colloquial or formal (for example the word *pretty* is far more likely to have the meaning "rather" in informal genres than in formal ones). In information retrieval, genre classification could enable users to sort search results according to their immediate interests. People who go into a bookstore or library are not usually looking simply for information about a particular topic, but rather have requirements of genre as well: they are looking for scholarly articles about hypnotism, novels about the French Revolution, editorials about the supercollider, and so forth.

If genre classification is so useful, why hasn't it figured much in computational linguistics before now? One important reason is that, up to now, the digitized corpora and collections which are the subject of much CL research have been for the most part generically homogeneous (i.e., collections of scientific abstracts or newspaper articles, encyclopedias, and so on), so that the problem of genre identification could be set aside. To a large extent, the problems of genre classification don't become salient until we are confronted with large and heterogeneous search domains like the World-Wide Web.

Another reason for the neglect of genre, though, is that it can be a difficult notion to get a conceptual handle on, particularly in contrast with properties of structure or topicality, which for all their complications involve well-explored territory. In order to do systematic work on automatic genre classification, by contrast, we require the answers to some basic theoretical and methodological questions. Is genre a single property or attribute that can be neatly laid out in some hierarchical structure? Or are we really talking about a multidimensional space of properties that have little more in common than that they are more or less orthogonal to topicality? And once we have the theoretical prerequisites in place, we have to ask whether genre can be reliably identified by means of computationally tractable cues.

In a broad sense, the word "genre" is merely a literary substitute for "kind of text," and discussions of literary classification stretch back to Aris-

totle. We will use the term "genre" here to refer to any widely recognized class of texts defined by some common communicative purpose or other functional traits, provided the function is connected to some formal cues or commonalities and that the class is extensible. For example an editorial is a shortish prose argument expressing an opinion on some matter of immediate public concern, typically written in an impersonal and relatively formal style in which the author is denoted by the pronoun *we*. But we would probably not use the term "genre" to describe merely the class of texts that have the objective of persuading someone to do something, since that class — which would include editorials, sermons, prayers, advertisements, and so forth — has no distinguishing formal properties. At the other end of the scale, we would probably not use "genre" to describe the class of sermons by John Donne, since that class, while it has distinctive formal characteristics, is not extensible. Nothing hangs in the balance on this definition, but it seems to accord reasonably well with ordinary usage.

The traditional literature on genre is rich with classificatory schemes and systems, some of which might in retrospect be analyzed as simple attribute systems. (For general discussions of literary theories of genre, see, e.g., Butcher (1932), Dubrow (1982), Fowler (1982), Frye (1957), Hernadi (1972), Hobbes (1908), Staiger (1959), and Todorov (1978).) We will refer here to the attributes used in classifying genres as GENERIC FACETS. A facet is simply a property which distinguishes a class of texts that answers to certain practical interests, and which is moreover associated with a characteristic set of computable structural or linguistic properties, whether categorical or statistical, which we will describe as "generic cues." In principle, a given text can be described in terms of an indefinitely large number of facets. For example, a newspaper story about a Balkan peace initiative is an example of a BROADCAST as opposed to DIRECTED communication, a property that correlates formally with certain uses of the pronoun *you*. It is also an example of a NARRATIVE, as opposed to a DIRECTIVE (e.g., in a manual), SUASIVE (as in an editorial), or DESCRIPTIVE (as in a market survey) communication; and this facet correlates, among other things, with a high incidence of preterite verb forms.

Apart from giving us a theoretical framework for understanding genres, facets offer two practical advantages. First, some applications benefit from categorization according to facet, not genre. For example, in an information retrieval context, we will want to consider the OPINION feature most highly when we are searching for public reactions to the supercollider, where newspaper columns, editorials, and letters to the editor will be of roughly equal interest. For other purposes we will want to stress narrativity, for example in looking for accounts of the storming of the Bastille in either novels or histories.

Secondly, we can extend our classification to genres not previously encountered. Suppose that we are presented with the unfamiliar category FINANCIAL ANALYSTS' REPORT. By analyzing genres as bundles of facets, we can categorize this genre as INSTITUTIONAL (because of the use of *we* as in editorials and annual reports) and as NON-SUASIVE or non-argumentative (because of the low incidence of question marks, among other things), whereas a system trained on genres as atomic entities would not be able to make sense of an unfamiliar category.

1.1 Previous Work on Genre Identification

The first linguistic research on genre that uses quantitative methods is that of Biber (1986; 1988; 1992; 1995), which draws on work on stylistic analysis, readability indexing, and differences between spoken and written language. Biber ranks genres along several textual "dimensions", which are constructed by applying factor analysis to a set of linguistic syntactic and lexical features. Those dimensions are then characterized in terms such as "informative vs. involved" or "narrative vs. non-narrative." Factors are not used for genre classification (the values of a text on the various dimensions are often not informative with respect to genre). Rather, factors are used to validate hypotheses about the functions of various linguistic features.

An important and more relevant set of experiments, which deserves careful attention, is presented in Karlgren and Cutting (1994). They too begin with a corpus of hand-classified texts, the Brown corpus. One difficulty here, however, is that it is not clear to what extent the Brown corpus classification used in this work is relevant for practical or theoretical purposes. For example, the category "Popular Lore" contains an article by the decidedly highbrow Harold Rosenberg from *Commentary,* and articles from *Model Railroader* and *Gourmet,* surely not a natural class by any reasonable standard. In addition, many of the text features in Karlgren and Cutting are structural cues that require tagging. We will replace these cues with two new classes of cues that are easily computable: character-level cues and deviation cues.

2 Identifying Genres: Generic Cues

This section discusses generic cues, the "observable" properties of a text that are associated with facets.

2.1 Structural Cues

Examples of structural cues are passives, nominalizations, topicalized sentences, and counts of the frequency of syntactic categories (e.g., part-of-speech tags). These cues are not much discussed in the traditional literature on genre, but have come to the fore in recent work (Biber, 1995; Karlgren and Cutting, 1994). For purposes of automatic classification they have the limitation that they require tagged or parsed texts.

2.2 Lexical Cues

Most facets are correlated with lexical cues. Examples of ones that we use are terms of address (e.g., *Mr.*, *Ms.*), which predominate in papers like the *New York Times*; Latinate affixes, which signal certain highbrow registers like scientific articles or scholarly works; and words used in expressing dates, which are common in certain types of narrative such as news stories.

2.3 Character-Level Cues

Character-level cues are mainly punctuation cues and other separators and delimiters used to mark text categories like phrases, clauses, and sentences (Nunberg, 1990). Such features have not been used in previous work on genre recognition, but we believe they have an important role to play, being at once significant and very frequent. Examples include counts of question marks, exclamations marks, capitalized and hyphenated words, and acronyms.

2.4 Derivative Cues

Derivative cues are ratios and variation measures derived from measures of lexical and character-level features.

Ratios correlate in certain ways with genre, and have been widely used in previous work. We represent ratios implicitly as sums of other cues by transforming all counts into natural logarithms. For example, instead of estimating separate weights α, β, and γ for the ratios words per sentence (average sentence length), characters per word (average word length) and words per type (token/type ratio), respectively, we express this desired weighting:

$$\alpha \log \frac{W+1}{S+1} + \beta \log \frac{C+1}{W+1} + \gamma \log \frac{W+1}{T+1}$$

as follows:

$$(\alpha - \beta + \gamma)\log(W+1) - \alpha \log(S+1) + \beta \log(C+1) - \gamma \log(T+1)$$

(where W = word tokens, S = sentences, C = characters, T = word types). The 55 cues in our experiments can be combined to almost 3000 different ratios. The log representation ensures that all these ratios are available implicitly while avoiding overfitting and the high computational cost of training on a large set of cues.

Variation measures capture the amount of variation of a certain count cue in a text (e.g., the standard deviation in sentence length). This type of useful metric has not been used in previous work on genre.

The experiments in this paper are based on 55 cues from the last three groups: lexical, character-level and derivative cues. These cues are easily computable in contrast to the structural cues that have figured prominently in previous work on genre.

3 Method

3.1 Corpus

The corpus of texts used for this study was the Brown Corpus. For the reasons mentioned above, we used our own classification system, and eliminated texts that did not fall unequivocally into one of our categories. We ended up using 499 of the 802 texts in the Brown Corpus. (While the Corpus contains 500 *samples*, many of the samples contain several texts.)

For our experiments, we analyzed the texts in terms of three categorical facets: BROW, NARRATIVE, and GENRE. BROW characterizes a text in terms of the presumptions made with respect to the required intellectual background of the target audience. Its levels are POPULAR, MIDDLE, UPPER-MIDDLE, and HIGH. For example, the mainstream American press is classified as MIDDLE and tabloid newspapers as POPULAR. The NARRATIVE facet is binary, telling whether a text is written in a narrative mode, primarily relating a sequence of events. The GENRE facet has the values REPORTAGE, EDITORIAL, SCITECH, LEGAL, NONFICTION, FICTION. The first two characterize two types of articles from the daily or weekly press: reportage and editorials. The level SCITECH denominates scientific or technical writings, and LEGAL characterizes various types of writings about law and government administration. Finally, NONFICTION is a fairly diverse category encompassing most other types of expository writing, and FICTION is used for works of fiction.

Our corpus of 499 texts was divided into a training subcorpus (402 texts) and an evaluation subcorpus (97). The evaluation subcorpus was designed

to have approximately equal numbers of all represented combinations of facet levels. Most such combinations have six texts in the evaluation corpus, but due to small numbers of some types of texts, some extant combinations are underrepresented. Within this stratified framework, texts were chosen by a pseudo random-number generator. This setup results in different quantitative compositions of training and evaluation set. For example, the most frequent genre level in the training subcorpus is REPORTAGE, but in the evaluation subcorpus NONFICTION predominates.

3.2 Logistic Regression

We chose logistic regression (LR) as our basic numerical method. Two informal pilot studies indicated that it gave better results than linear discrimination and linear regression.

LR is a statistical technique for modeling a binary response variable by a linear combination of one or more predictor variables, using a logit link function:

$$g(\pi) = log(\pi/(1-\pi))$$

and modeling variance with a binomial random variable, i.e., the dependent variable $log(\pi/(1-\pi))$ is modeled as a linear combination of the independent variables. The model has the form $g(\pi) = x_i\beta$ where π is the estimated response probability (in our case the probability of a particular facet value), x_i is the feature vector for text i, and β is the weight vector which is estimated from the matrix of feature vectors. The optimal value of β is derived via maximum likelihood estimation (McCullagh and Nelder, 1989), using SPlus (Statistical Sciences, 1991).

For binary decisions, the application of LR was straightforward. For the polytomous facets GENRE and BROW, we computed a predictor function independently for each level of each facet and chose the category with the highest prediction.

The most discriminating of the 55 variables were selected using stepwise backward selection based on the AIC criterion (see documentation for STEP.GLM in Statistical Sciences (1991)). A separate set of variables was selected for each binary discrimination task.

3.2.1 Structural Cues

In order to see whether our easily-computable surface cues are comparable in power to the structural cues used in Karlgren and Cutting (1994), we also ran LR with the cues used in their experiment. Because we use individual texts in our experiments instead of the fixed-length conglomerate samples of Karlgren and Cutting, we averaged all count features over text length.

3.3 Neural Networks

Because of the high number of variables in our experiments, there is a danger that overfitting occurs. LR also forces us to simulate polytomous decisions by a series of binary decisions, instead of directly modeling a multinomial response. Finally, classical LR does not model variable interactions.

For these reasons, we ran a second set of experiments with neural networks, which generally do well with a high number of variables because they protect against overfitting. Neural nets also naturally model variable interactions. We used two architectures, a simple perceptron (a two-layer feed-forward network with all input units connected to all output units), and a multi-layer perceptron with all input units connected to all units of the hidden layer, and all units of the hidden layer connected to all output units. For binary decisions, such as determining whether or not a text is NARRATIVE, the output layer consists of one sigmoidal output unit; for polytomous decisions, it consists of four (BROW) or six (GENRE) softmax units (which implement a multinomial response model) (Rumelhart et al., 1995). The size of the hidden layer was chosen to be three times as large as the size of the output layer (3 units for binary decisions, 12 units for BROW, 18 units for GENRE).

For binary decisions, the simple perceptron fits a logistic model just as LR does. However, it is less prone to overfitting because we train it using three-fold cross-validation. Variables are selected by summing the cross-entropy error over the three validation sets and eliminating the variable that if eliminated results in the lowest cross-entropy error. The elimination cycle is repeated until this summed cross-entropy error starts increasing. Because this selection technique is time-consuming, we only apply it to a subset of the discriminations.

4 Results

Table 1 gives the results of the experiments. For each genre facet, it compares our results using surface cues (both with logistic regression and neural nets) against results using Karlgren and Cutting's structural cues on the one hand (last pair of columns) and against a baseline on the other (first column). Each text in the evaluation suite was tested for each facet. Thus the number 78 for NARRATIVE under method "LR (Surf.) All" means that when all texts were subjected to the NARRATIVE test, 78% of them were classified correctly.

There are at least two major ways of conceiving what the baseline should be in this experiment. If

the machine were to guess randomly among k categories, the probability of a correct guess would be $1/k$, i.e., 1/2 for NARRATIVE, 1/6 for GENRE, and 1/4 for BROW. But one could get dramatic improvement just by building a machine that always guesses the most populated category: NONFICT for GENRE, MIDDLE for BROW, and No for NARRATIVE. The first approach would be fair, because our machines in fact have no prior knowledge of the distribution of genre facets in the evaluation suite, but we decided to be conservative and evaluate our methods against the latter baseline. No matter which approach one takes, however, each of the numbers in the table is significant at $p < .05$ by a binomial distribution. That is, there is less than a 5% chance that a machine guessing randomly could have come up with results so much better than the baseline.

It will be recalled that in the LR models, the facets with more than two levels were computed by means of binary decision machines for each level, then choosing the level with the most positive score. Therefore some feeling for the internal functioning of our algorithms can be obtained by seeing what the performance is for each of these binary machines, and for the sake of comparison this information is also given for some of the neural net models. Table 2 shows how often each of the binary machines correctly determined whether a text did or did not fall in a particular facet level. Here again the appropriate baseline could be determined two ways. In a machine that chooses randomly, performance would be 50%, and all of the numbers in the table would be significantly better than chance ($p < .05$, binomial distribution). But a simple machine that always guesses No would perform much better, and it is against this stricter standard that we computed the baseline in Table 2. Here, the binomial distribution shows that some numbers are not significantly better than the baseline. The numbers that are significantly better than chance at $p < .05$ by the binomial distribution are starred.

Tables 1 and 2 present aggregate results, when all texts are classified for each facet or level. Table 3, by contrast, shows which classifications are assigned for texts that actually belong to a specific known level. For example, the first row shows that of the 18 texts that really are of the REPORTAGE GENRE level, 83% were correctly classified as REPORTAGE, 6% were misclassified as EDITORIAL, and 11% as NONFICTION. Because of space constraints, we present this amount of detail only for the six GENRE levels, with logistic regression on selected surface variables.

5 Discussion

The experiments indicate that categorization decisions can be made with reasonable accuracy on the basis of surface cues. All of the facet level assignments are significantly better than a baseline of always choosing the most frequent level (Table 1), and the performance appears even better when one considers that the machines do not actually know what the most frequent level is.

When one takes a closer look at the performance of the component machines, it is clear that some facet levels are detected better than others. Table 2 shows that within the facet GENRE, our systems do a particularly good job on REPORTAGE and FICTION, trend correctly but not necessarily significantly for SCITECH and NONFICTION, but perform less well for EDITORIAL and LEGAL texts. We suspect that the indifferent performance in SCITECH and LEGAL texts may simply reflect the fact that these genre levels are fairly infrequent in the Brown corpus and hence in our training set. Table 3 sheds some light on the other cases. The lower performance on the EDITORIAL and NONFICTION tests stems mostly from misclassifying many NONFICTION texts as EDITORIAL. Such confusion suggests that these genre types are closely related to each other, as in fact they are. Editorials might best be treated in future experiments as a subtype of NONFICTION, perhaps distinguished by separate facets such as OPINION and INSTITUTIONAL AUTHORSHIP.

Although Table 1 shows that our methods predict BROW at above-baseline levels, further analysis (Table 2) indicates that most of this performance comes from accuracy in deciding whether or not a text is HIGH BROW. The other levels are identified at near baseline performance. This suggests problems with the labeling of the BROW feature in the training data. In particular, we had labeled journalistic texts on the basis of the overall brow of the host publication, a simplification that ignores variation among authors and the practice of printing features from other publications. We plan to improve those labelings in future experiments by classifying brow on an article-by-article basis.

The experiments suggest that there is only a small difference between surface and structural cues. Comparing LR with surface cues and LR with structural cues as input, we find that they yield about the same performance: averages of 77.0% (surface) vs. 77.5% (structural) for all variables and 78.4% (surface) vs. 78.9% (structural) for selected variables. Looking at the independent binary decisions on a task-by-task basis, surface cues are worse in 10 cases

Table 1: Classification Results for All Facets.

	Baseline	LR (Surf.)		2LP		3LP		LR (Struct.)	
Facet		All	Sel.	All	Sel.	All	Sel.	All	Sel.
Narrative	54	78	80	82	82	86	82	78	80
Genre	33	61	66	75	79	71	74	66	62
Brow	32	44	46	47	—	54	—	46	53

Note. Numbers are the percentage of the evaluation subcorpus ($N = 97$) which were correctly assigned to the appropriate facet level; the Baseline column tells what percentage would be correct if the machine always guessed the most frequent level. LR is Logistic Regression, over our surface cues (Surf.) or Karlgren and Cutting's structural cues (Struct.); 2LP and 3LP are 2- or 3-layer perceptrons using our surface cues. Under each experiment, All tells the results when all cues are used, and Sel. tells the results when for each level one selects the most discriminating cues. A dash indicates that an experiment was not run.

Table 2: Classification Results for Each Facet Level.

	Baseline	LR (Surf.)		2LP	3LP	LR (Struct.)	
Levels		All	Sel.	All	All	All	Sel.
Genre							
Rep	81	89*	88	94*	94*	90*	90*
Edit	81	75	75	74	80	79	77
Legal	95	96	96	95	95	93	93
Scitech	94	100*	96	99*	94	93	96
Nonfict	67	67	68	78*	67	73	74
Fict	81	93*	96*	99*	81	96*	96*
Brow							
Popular	74	74	75	74	74	72	73
Middle	68	66	67	64	64	58	64
Uppermiddle	88	74	78	86	88	79	82
High	70	84*	88*	89*	90*	85*	86*

Note. Numbers are the percentage of the evaluation subcorpus ($N = 97$) which was correctly classified on a binary discrimination task. The Baseline column tells what percentage would be got correct by guessing No for each level. Headers have the same meaning as in Table 1.
* means significantly better than Baseline at $p < .05$, using a binomial distribution (N=97, p as per first column).

Table 3: Genre Binary Level Classification Results by Genre Level.

Actual	Guess						
	Rep	Edit	Legal	Scitech	Nonfict	Fict	N
Rep	83	6	0	0	11	0	18
Edit	17	61	0	0	17	6	18
Legal	20	0	20	0	60	0	5
Scitech	0	0	0	83	17	0	6
Nonfict	3	34	0	6	47	9	32
Fict	0	6	0	0	0	94	18

Note. Numbers are the percentage of the texts actually belonging to the GENRE level indicated in the first column that were classified as belonging to each of the GENRE levels indicated in the column headers. Thus the diagonals are correct guesses, and each row would sum to 100%, but for rounding error.

and better in 8 cases. Such a result is expected if we assume that either cue representation is equally likely to do better than the other (assuming a binomial model, the probability of getting this or a more extreme result is $\sum_{i=0}^{8} b(i; 18, 0.5) = 0.41$). We conclude that there is at best a marginal advantage to using structural cues, an advantage that will not justify the additional computational cost in most cases.

Our goal in this paper has been to prepare the ground for using genre in a wide variety of areas in natural language processing. The main remaining technical challenge is to find an effective strategy for variable selection in order to avoid overfitting during training. The fact that the neural networks have a higher performance on average and a much higher performance for some discriminations (though at the price of higher variability of performance) indicates that overfitting and variable interactions are important problems to tackle.

On the theoretical side, we have developed a taxonomy of genres and facets. Genres are considered to be generally reducible to bundles of facets, though sometimes with some irreducible atomic residue. This way of looking at the problem allows us to define the relationships between different genres instead of regarding them as atomic entities. We also have a framework for accommodating new genres as yet unseen bundles of facets. Finally, by decomposing genres into facets, we can concentrate on whatever generic aspect is important in a particular application (e.g., narrativity for one looking for accounts of the storming of the Bastille).

Further practical tests of our theory will come in applications of genre classification to tagging, summarization, and other tasks in computational linguistics. We are particularly interested in applications to information retrieval where users are often looking for texts with particular, quite narrow generic properties: authoritatively written documents, opinion pieces, scientific articles, and so on. Sorting search results according to genre will gain importance as the typical data base becomes increasingly heterogeneous. We hope to show that the usefulness of retrieval tools can be dramatically improved if genre is one of the selection criteria that users can exploit.

References

Biber, Douglas. 1986. Spoken and written textual dimensions in English: Resolving the contradictory findings. *Language*, 62(2):384–413.

Biber, Douglas. 1988. *Variation across Speech and Writing*. Cambridge University Press, Cambridge, England.

Biber, Douglas. 1992. The multidimensional approach to linguistic analyses of genre variation: An overview of methodology and finding. *Computers in the Humanities*, 26(5–6):331–347.

Biber, Douglas. 1995. *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge University Press, Cambridge, England.

Butcher, S. H., editor. 1932. *Aristotle's Theory of Poetry and Fine Arts, with The Poetics*. Macmillan, London. 4th edition.

Dubrow, Heather. 1982. *Genre*. Methuen, London and New York.

Fowler, Alistair. 1982. *Kinds of Literature*. Harvard University Press, Cambridge, Massachusetts.

Frye, Northrop. 1957. *The Anatomy of Criticism*. Princeton University Press, Princeton, New Jersey.

Hernadi, Paul. 1972. *Beyond Genre*. Cornell University Press, Ithaca, New York.

Hobbes, Thomas. 1908. The answer of mr Hobbes to Sir William Davenant's preface before Gondibert. In J.E. Spigarn, editor, *Critical Essays of the Seventeenth Century*. The Clarendon Press, Oxford.

Karlgren, Jussi and Douglass Cutting. 1994. Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of Coling 94*, Kyoto.

McCullagh, P. and J.A. Nelder. 1989. *Generalized Linear Models*, chapter 4, pages 101–123. Chapman and Hall, 2nd edition.

Nunberg, Geoffrey. 1990. *The Linguistics of Punctuation*. CSLI Publications, Stanford, California.

Rumelhart, David E., Richard Durbin, Richard Golden, and Yves Chauvin. 1995. Backpropagation: The basic theory. In Yves Chauvin and David E. Rumelhart, editors, *Backpropagation: Theory, Architectures, and Applications*. Lawrence Erlbaum, Hillsdale, New Jersey, pages 1–34.

Staiger, Emil. 1959. *Grundbegriffe der Poetik*. Atlantis, Zurich.

Statistical Sciences. 1991. *S-PLUS Reference Manual*. Statistical Sciences, Seattle, Washington.

Todorov, Tsvetan. 1978. *Les genres du discours*. Seuil, Paris.