Research Report

THE VOCABULARIES OF ACADEMIA

Stanley Schachter,[1] Frances Rauscher,[1] Nicholas Christenfeld,[2] and Kimberly Tyson Crone[1]

[1]*Columbia University and* [2]*University of California, San Diego*

Abstract—*It has been demonstrated that humanists are far more likely to use filled pauses ("uh," "ah," or "um") during their lectures than are social or natural scientists This finding has been interpreted in terms of the hypothesis that filled pauses indicate time out while the speaker searches for the next word or phrase Based on the assumption that the more options at a choice point, the more likely a speaker will say "uh," it is hypothesized that the humanities are characterized by richer vocabularies (i e , more synonyms) than are the sciences An analysis of the number of different words used in lectures and in professional publications indicates that this is indeed the case Scientists consistently use fewer different words than do humanists Further, the number of different words correlates positively with the frequency of saying "uh" during lectures These findings are not restricted to academics, for in newspaper accounts, journalists use fewer different words in stories about science than in stories about the arts*

It has been demonstrated (Schachter, Christenfeld, Ravina, & Bilous, 1991) that humanists, in their lectures, are far more likely to use filled pauses, that is, to say "uh," "ah," "er," and the like, than are either social or natural scientists This difference seems to be due to the subject matter proper rather than to differences in fluency or to differential personality characteristics of the types of people attracted to the various academic disciplines, for when speaking on a common, nondisciplinary topic, these same lecturers are virtually identical in their use of filled pauses Nor can these differences be explained by any of a variety of possible artifacts such as age, sex, teaching experience, or preparation time, for the humanists and scientists observed were similar in all of these respects

In an attempt to explain these differences in the use of filled pauses, we (Schachter et al , 1991) relied on the hypothesis that pauses in speech are indications of time out while the speech production apparatus searches for the next word, phrase, or idea (Goldman-Eisler, 1968, Rochester, 1973) It should follow that the more options at a choice point in speech, the greater the likelihood that the speaker will pause

It is suggested that, as related to options, the sciences and the humanities differ in the following respects

- First, the scientific method proper imposes strict limits on the options Given a clearly stated set of assumptions, the derivations must follow The social sciences aspire to this model The humanities are largely expositional
- Next, the sciences, basic or applied, are eventually concerned with facts and, again, there are few, if any, options Under precisely specified conditions, the freezing point of helium is such and such—and that is it
- Finally, we have suggested that there are simply fewer synonyms for the basic terms of the natural sciences than for the terms of either the social sciences or the humanities There are, for example, no synonyms for *molecule* or *atom* or *ion Salt* may serve as a substitute for *sodium chloride* or *NaCl*, but that is about it In contrast, consider the alternatives for *love, beauty, group structure, prejudice,* or *style*

It is the purpose of this report to test directly the implications of these notions about synonyms and the various academic disciplines If it is correct that the sciences have fewer synonyms for the basic terms and constructs of their subject matter than do the humanities, it should be expected that natural scientists will use fewer different terms in discoursing upon their subject than will humanists We present data testing this hypothesis in lectures, interviews, technical articles, Ph D theses, and journalistic popularizations

METHOD

The technique for collecting data on synonyms was simple Using a Macintosh spread sheet program, we analyzed the first 400 words of any scholarly product—undergraduate lecture, article, or Ph D thesis—for the number of different words used—a method of analysis sometimes called a type-token ratio analysis Variations in tense, case, number, and so on were ignored so that, for example, "bifurcation" and "bifurcations" would count as the same word, as would "buy" and "bought" or "is," "was," and "were " For each lecture, then, a single number was calculated the number of unique words out of the first 400 words spoken With training, this analysis could be carried out accurately with an intraclass reliability coefficient of R = 99 between independent raters

In the original study of filled pauses, the data were collected by on-the-spot observation of lecturers as well as, whenever possible, by tape-recording the lectures These transcribed tapes provided the data for the present analyses For one reason or another, it was not always possible to make satisfactory transcriptions, and the number of cases in the tables in this report do not fully correspond to the number of cases in similar tables in the study of filled pauses (Schachter et al , 1991) To make up for some of these omissions, we transcribed three additional undergraduate lectures, one in chemistry and two in psychology We note immediately that, regarding the

Frances Rauscher is presently a Research Fellow of the Center for the Neurobiology of Learning and Memory at the University of California, Irvine

Address correspondence to Stanley Schachter, Department of Psychology, Columbia University, New York, NY 10027

analysis of filled pauses, these lecturers were completely typical of their field, in all cases falling close to the mean of their discipline

On this same point, we note that, throughout this report, the number of cases varies slightly from table to table This is the unfortunate but probably inevitable consequence of the fact that we were trying to assess the verbal behavior of the same group of people during lectures, during interviews, and in professional publications Not only did the recording apparatus occasionally fail, but also a few people left town before they could be interviewed, some of the lecturers had never published, and so on Except for Table 4, all tables include all of the data available on the point at issue In Table 4, we analyze the data just for those subjects whose data record is complete (i e , we have understandable transcriptions of their lectures and interviews, and they have published an article in one of their discipline's journals)

RESULTS

The distribution of lectures analyzed and the mean number of different words used out of the first 400[1] are presented in Table 1 There are substantial differences in the sizes of the vocabularies in the various departments, $F(9, 24) = 4\ 16, p < 005$, with the natural scientists using the fewest different words, and the humanists the most [2] That these differences in vocabulary may, to some extent, be causally related to the frequency of filled pauses is suggested by the fact that for these 34 lecturers, the correlation[3] between the number of different words and the frequency of "uh" per minute during the lecture is $r = + 34, p < 05$

Table 1 *Number of different words out of the first 400 words spoken in introductory lectures*

Discipline	*N*	Number of different words
Natural sciences		
Biology	4	158 5
Chemistry	4	151 3
Mathematics	4	138 5
Psychology	2	175 0
Total	14	153 1
Social sciences		
Economics	3	144 0
Political science	4	167 8
Sociology	3	151 7
Total	10	155 8
Humanities		
Art history	5	196 2
English	3	192 0
Philosophy	2	167 5
Total	10	189 2

It should be noted that in the study of filled pauses, the one exception to the humanists' high "um" rate (i e , frequency of filled pauses) was philosophy Philosophy's faculty used "um" like scientists As can be seen in Table 1, they also use words as scientists do, with a far smaller vocabulary than the art historians and English professors with whom they are grouped

This vocabulary analysis does lend support to the hypothesis that the pattern of filled-pause rates in academic departments reflects the number of options that a speaker must ponder Of course, it is possible that the size of the vocabulary is not a function of the discipline itself, but rather that people with larger vocabularies are more likely to become professors of English than of chemistry Or, conceivably, continued exposure to a particular field may change the size of a professor's vocabulary If either of these possibilities were the case, then one would expect that humanists would use larger vocabularies than scientists even when they all were speaking on the same subject

In the previous work on filled pauses, just such speech samples were obtained, when the same faculty members were interviewed on a common topic—graduate training The fact that their "um" rates were identical when the subject matter was identical suggested that subject matter, rather than speaker, was the important factor If "um" rate does reflect vocabulary size, when speaking on a common topic, people from the various departments should all have vocabularies of the same size

Vocabulary analyses were performed on the first 400 words of the interviews, and the means are presented in Table 2 Clearly, when the topic is the same, the differences between scientists and humanists vanish, $F(2, 33) = 0\ 19$, n s This result offers further support for the notion that there is something about giving a science lecture, rather than being a

1 In the case of class lectures, the word count began with the formal lecture proper and did not include any preliminary announcements about exams, office hours, and the like

2 In order to convey some feeling for the nature of these data, we offer the following two typical passages extracted from our lecture transcripts The first quotation was taken from the lecture of an art history professor and typifies the diversity of vocabulary used by humanists

" we notice how really different these lyric contemporary works made for an exposition both for the same constituent are in the style and in the dress um the um dynamic um the well the theatrical epiphany of the triumphant general and emperor Phidus here this panel in the arched hippodrome sticking in from "

There are a total of 50 words in this selection of which 38 are different In contrast, the following 50-word passage was taken from a biologist's lecture

that the introns contain genes themselves in fact in one case of a very large gene with very large introns one of the introns contained a gene that makes premessage that is itself spliced so you have a gene within the coding or within the DNA domain of another gene

Of the 50 words in this selection, 29 are different It should also be noticed that, all told, the biologist uses a greater number of nouns and repeats these nouns far more often than does the art historian This pattern appears to be generally characteristic of the sciences versus the humanities An overall analysis of the grammatical differences in language use of the sciences and arts is under way

3 Obviously, the implication of causality must be tempered by the usual caveats about correlation coefficients in field studies such as this one Perhaps, as we suggest, vocabulary size is causally related to the production of filled pauses, perhaps they are both mediated by some third factor, such as Tetlock's (1983) "cognitive complexity," from which conceivably both filled pauses and vocabulary richness could be independently derived Though the facts about pauses and vocabulary size seem well established, interpretation will require further work

Table 2 *Number of different words out of the first 400 words spoken in interviews*

Discipline	*N*	Number of different words
Natural sciences		
Biology	3	167 0
Chemistry	4	167 8
Mathematics	4	171 5
Psychology	4	163 0
Total	15	167 3
Social sciences		
Economics	3	172 7
Political science	4	164 8
Sociology	3	165 7
Total	10	167 4
Humanities		
Art history	5	168 2
English	3	175 7
Philosophy	3	167 7
Total	11	170 1

science lecturer, that reduces the number of different words used

We have, then, two findings First, giving a lecture in the humanities confronts the speaker with more options than giving a lecture in the sciences and is accompanied by greater use of filled pauses Second, when speaking on a nondisciplinary, uniform topic, a scientist faces the same number of word options as a humanist and says "um" just as often Other explanations are, of course, possible, but these results do, at least, suggest that choosing from a greater number of words may lead to a greater rate of filled pauses

If this reasoning is correct, then one should expect that the vocabulary difference would not be confined to undergraduate lectures It should also be manifest in the professional writing of the disciplines To check this hypothesis, we performed the same analyses on the most recent publication of as many of these lecturers as had published within the past 10 years If the smaller vocabulary of the sciences is simply some sort of pedagogical concession to novices, then it should not exist when the scientists are writing for sophisticated colleagues If, however, the effect is something more intrinsic to the structure of knowledge of the field, then the vocabulary should again be smaller for the scientists

We attempted to track down the most recent publication in a professional journal, as identified by an on-line library search service, for each faculty member whose lecture had been coded Some of the younger faculty had not yet published, and they were excluded from the analyses The data for the vocabulary used in the first 400 words of published articles are presented in Table 3 Mathematicians have been excluded because they appear not to use words as we know them In most disciplines, papers at least start with traditional prose, but a paper in mathematics is incomprehensible to the outsider from the very beginning Rather than try to determine the analogs for English words in the string of formulas that is a mathematics article, we chose to ignore the field In any event, two clear findings emerge First, the pattern of vocabulary sizes is replicated, $F(2, 33) = 6\ 67$, $p < 004$ Natural scientists use the fewest different words in their articles, and humanists use the most Again, philosophers are lower than their humanist standing would predict, and are more in line with the scientists Second, in almost all departments,

Table 3 *Number of different words out of the first 400 words written in professional publications*

Discipline	*N*	Number of different words
Natural sciences		
Biology	4	178 3
Chemistry	5	177 4
Psychology	5	163 0
Total	14	172 5
Social sciences		
Economics	4	189 0
Political science	4	200 5
Sociology	2	172 5
Total	10	190 3
Humanities		
Art history	5	215 0
English	4	200 0
Philosophy	3	175 0
Total	12	200 0

there are more different words used when writing than when speaking, a marked effect ($F[1, 68] = 14\ 99$, $p < 0002$) due in part, undoubtedly, to the anathema with which copyeditors seem to regard redundancy and repetition

Finally, we note that the correlation between "um" rate in the lectures of these 36 academics and the number of different words in their publications is $r = + 44$, $p < 01$ It is surprising that this is somewhat larger than the correlation between the number of different words in lectures and the "um" rate ($r = + 34$) We suspect that this is an indication that the lexical idiosyncrasies of a discipline are more evident in a technical publication than in an undergraduate lecture Technical publications stick strictly to the subject, but it is a rare lecturer who can resist the jokes, asides, and stories that have little to do with the subject but are presumed to be the mark of an "interesting" or, at least, an "entertaining" lecturer

In order to control for any possible artifactual effects of the fact that the *N* varies among the previous tables, we include in Table 4 only the findings for those subjects for whom we have data on vocabulary size in all three communication modalities lectures, interviews, and articles It is immediately clear that the previous findings are strongly supported Controlling for individual differences, humanists use many more different words in both lectures and articles than do either natural or social scientists In their interviews, the three groups are virtually identical

We turn next to one more comparison of the vocabularies of the various disciplines—in doctoral dissertations We simply went to the library and found the most recent dissertations on file in each of the fields The results of these analyses are presented in Table 5 Again, mathematics has been excluded, because by the time they receive their doctorates, its students have already learned not to use English These data suggest that by the time of receiving a Ph D , a student has already learned the vocabulary characteristic of his or her field The trend is the same as for the faculty, with the humanists using more words than the scientists, and the social scientists falling in between, $F(2, 27) = 3\ 41$, $p < 05$

Next, we ask whether these differ-

Vocabularies of Academia

Table 4. *Number of different words out of the first 400 used by the same persons in lectures, interviews, and publications*

Discipline	N	Lectures	Interviews	Publications
Natural sciences				
Biology	2	157 0	158 5	170 0
Chemistry	4	151 3	167 8	170 5
Psychology	2	175 0	167 5	152 0
Total	8	158 6	165 4	165 8
Social sciences				
Economics	2	142 5	166 5	170 5
Political science	4	167 8	164 8	200 5
Sociology	2	156 0	159 5	172 5
Total	8	158 5	163 9	186 0
Humanities				
Art history	5	196 2	168 2	215 0
English	2	198 0	176 5	198 0
Philosophy	2	167 5	169 0	169 0
Total	9	190 2	170 2	201 0
	F	4 985	0 384	6 568
	df	8, 16	8, 16	8, 16
	p	< 004	n s	< 001

ences among the disciplines are maintained in journalistic popularizations of scientific or artistic work, that is, write-ups, usually by nonprofessionals of professional work, that are intended to be read by nonprofessionals One can argue cogently either way On the one hand, there is no synonym for *atom* or *molecule*, whether one is writing for a professional or a lay audience On the other hand, one might guess that the journalist's need to explain technical terms would lead to the use of a great variety of new and different words Rather than debate the issue, let us turn to the results of an analysis of stories published in the *New York Times* between July 2, 1991, and October 24, 1991 We analyzed 12 stories on discoveries in the sciences, including biology, chemistry, physics, and astronomy We also analyzed 12 stories in the arts, including stories concerned with major new exhibitions and concerts and new discoveries about the lives and work of major poets, painters, and the like These were all major stories, and we simply counted the number of different words in the first 400 words of each story The sciences averaged 204 3 different words per story, and the arts 218 9 different words, $F(1, 22) = 6\ 94, p < 02$ It does appear, then, that the nature of the subject matter determines the number of different words used in public presentations, whether undergraduate lectures, technical publications, theses, or journalistic accounts

Table 5 *Number of different words out of the first 400 words written in dissertations*

Discipline	N	Number of different words
Natural sciences		
Biology	2	176 0
Chemistry	3	174 3
Psychology	6	171 2
Total	11	172 9
Social sciences		
Economics	2	152 0
Political science	4	194 3
Sociology	3	193 3
Total	9	184 6
Humanities		
Art history	3	202 0
English	4	210 0
Philosophy	3	175 7
Total	10	197 3

The fact that a difference in vocabulary size between the sciences and the arts is maintained in journalistic accounts suggests that there is something inherent in these academic enterprises that determines vocabulary size The fact that these differences are maintained in newspaper stories opens the possibility of testing these ideas about vocabulary size in domains other than the purely academic After all, science is hardly the only human enterprise with a succinct, terse vocabulary and a precise, mandated set of rules Sports and games such as bridge or backgammon are at least as constrained in these respects as even the most rigorous of the sciences The rules of baseball, for example, are quite as precise and their implications quite as clear-cut as virtually any theoretical construction in science There may be no synonyms for *ion* or *atom* or *molecule*, but there are also no synonyms for *strike* or *ball* or *out*, and though one occasionally speaks of a *three-bagger* or a *three-base hit*, these are about the only synonyms that come to mind for a *triple* All of this suggests, of course, that the vocabulary size for reports of games and sport events may be just as limited as for reports of discoveries in the sciences

Similarly, lecturing or writing about the arts is hardly the only human endeavor that relies on interpretation, evaluation, and speculation Intuitively, one can think of editorial writing as characterized by similar qualities, and it seems a reasonable guess that editorials might rely on at least as diverse a vocabulary as do stories on the arts

In order to learn if there was anything to these speculations, in addition to analyzing the vocabulary of stories about the arts and sciences in the *New York Times*, we analyzed lead editorials, bridge columns, and a variety of stories about sporting events (baseball, football, hockey, golf, and tennis) All these items appeared in the *Times* between August 12, 1991, and October 24, 1991 Other than that a story or an editorial had to be at least 400 words long,[4] there was no particular criterion for choosing the stories analyzed We simply chose whatever stories appeared in the paper on the days we were doing these analyses

4 Because bridge columns were rarely 400 words long, we adjusted the bridge data by calculating the percentage of different words and multiplying by 400

Table 6 *Number of different words out of the first 400 words written in editorials and art, science, sports, and bridge articles in the* New York Times

Type of story	*N*	Number of different words
Editorial	9	226 9
Art	12	218 9
Science	12	204 3
Sports	12	202 5
Bridge	17	205 0

Note $F(4, 57) = 6\ 57$, $p < 001$

The results of these analyses are presented in Table 6 Sports, bridge, and science are indeed virtually identical in the number of different words used, while both editorials and the arts are considerably higher than any of the other categories of stories Using Fisher's *t* test, editorials and arts do not differ significantly from one another, and each of them is significantly greater, with $p <$ 01, than either science, sports, or bridge articles Science, sports, and bridge do not differ significantly from one another

It does appear that this way of thinking about differences in vocabulary size goes beyond the academic disciplines and can probably be generalized to most human endeavors The sizes of the vocabularies of academicians and journalists follow the same trend When people are lecturing or writing about the sciences, sporting events, or bridge games, all of which have succinct vocabularies and precisely specified rules, they use fewer different words than do art and editorial writers and lecturers The significant correlations between "um" rate in lectures and vocabulary size in both lectures and publications suggest that people use more filled pauses when their vocabularies are not restricted That this relationship between filled pauses and vocabulary size holds true in areas other than the academic disciplines remains to be tested

REFERENCES

Goldman-Eisler F (1968) *Psycholinguistics Experiments in spontaneous speech* London Academic Press

Rochester S R (1973) The significance of pauses in spontaneous speech *Journal of Psycholinguistic Research 2* 51–81

Schachter S Christenfeld N J S Ravina B , & Bilous F R (1991) Speech disfluency and the structure of knowledge *Journal of Personality and Social Psychology 60* 362–367

Tetlock P E (1983) Accountability and complexity of thought *Journal of Personality and Social Psychology 45* 74–83

(RECEIVED 3/29/93, REVISION ACCEPTED 6/2/93)