AMERICAN
Scientist

FEATURE ARTICLE

The Science of Scientific Writing

If the reader is to grasp what the writer means, the writer must understand what the reader needs

George Gopen, Judith Swan

This article was originally published in the November-December 1990 issue of American Scientist.

Science is often hard to read. Most people assume that its difficulties are born out of necessity, out of the extreme complexity of scientific concepts, data and analysis. We argue here that complexity of thought need not lead to impenetrability of expression; we demonstrate a number of rhetorical principles that can produce clarity in communication without oversimplifying scientific issues. The results are substantive, not merely cosmetic: Improving the quality of writing actually improves the quality of thought.

The fundamental purpose of scientific discourse is not the mere presentation of information and thought, but rather its actual communication. It does not matter how pleased an author might be to have converted all the right data into sentences and paragraphs; it matters only whether a large majority of the reading audience accurately perceives what the author had in mind. Therefore, in order to understand how best to improve writing, we would do well to understand better how readers go about reading. Such an understanding has recently become available through work done in the fields of rhetoric, linguistics and cognitive psychology. It has helped to produce a methodology based on the concept of reader expectations.

Writing with the Reader in Mind: Expectation and Context

Readers do not simply read; they interpret. Any piece of prose, no matter how short, may "mean" in 10 (or more) different ways to 10 different readers. This methodology of reader expectations is founded on the recognition that readers make many of their most important interpretive decisions about the substance of prose based on clues they receive from its structure.

This interplay between substance and structure can be demonstrated by something as basic as a simple table. Let us say that in tracking the temperature of a liquid over a period of time, an investigator takes measurements every three minutes and records a list of temperatures. Those data could be presented by a number of written structures. Here are two possibilities:

t(time)=15', T(temperature)=32º, t=0', T=25º; t=6', T=29º; t=3', T=27º; t=12', T=32º; t=9'; T=31º

time (min)	temperature(ºC)
0	25
3	27
6	29
9	31
12	32
15	32

Precisely the same information appears in both formats, yet most readers find the second easier to interpret. It may be that the very familiarity of the tabular structure makes it easier to use. But, more significantly, the structure of the second table provides the reader with an easily perceived context (time) in which the significant piece of information (temperature) can be interpreted. The contextual material appears on the left in a pattern that produces an expectation of regularity; the interesting results appear on the right in a less obvious pattern, the discovery of which is the point of the table.

If the two sides of this simple table are reversed, it becomes much harder to read.

temperature(ºC)	time (min)
25	0
27	3
29	6
31	9
32	12
32	15

Since we read from left to right, we prefer the context on the left, where it can more effectively familiarize the reader. We prefer the new, important information on the right, since its job is to intrigue the reader.

Information is interpreted more easily and more uniformly if it is placed where most readers expect to find it. These needs and expectations of readers affect the interpretation not only of tables and illustrations but also of prose itself. Readers have relatively fixed expectations about where in the structure of prose they will encounter particular items of its substance. If writers can become consciously aware of these locations, they can better control the degrees of recognition and emphasis a reader will give to the various pieces of information being presented. Good writers are intuitively aware of these expectations; that is why their prose has what we call "shape."

This underlying concept of reader expectation is perhaps most immediately evident at the level of the largest units of discourse. (A unit of discourse is defined as anything with a beginning and an end: a clause, a sentence, a section, an article, etc.) A research article, for example, is generally divided into recognizable sections, sometimes labeled Introduction, Experimental Methods, Results and Discussion. When the sections are confused—when too much experimental detail is found in the Results section, or when discussion and results intermingle—readers are often equally confused. In smaller units of discourse the functional divisions are not so explicitly labeled, but readers have definite expectations all the same, and they search for certain information in particular places. If these structural expectations are continually violated, readers are forced to divert energy from understanding the content of a passage to

unraveling its structure. As the complexity of the context increases moderately, the possibility of misinterpretation or noninterpretation increases dramatically.

We present here some results of applying this methodology to research reports in the scientific literature. We have taken several passages from research articles (either published or accepted for publication) and have suggested ways of rewriting them by applying principles derived from the study of reader expectations. We have not sought to transform the passages into "plain English" for the use of the general public; we have neither decreased the jargon nor diluted the science. We have striven not for simplification but for clarification.

Reader Expectations for the Structure of Prose

Here is our first example of scientific prose, in its original form:

> The smallest of the URF's (URFA6L), a 207-nucleotide (nt) reading frame overlapping out of phase the NH_2-terminal portion of the
>
> adenosinetriphosphatase (ATPase) subunit 6 gene has been identified as the animal equivalent of the recently discovered yeast H^+-ATPase subunit 8 gene. The functional significance of the other URF's has been, on the contrary, elusive. Recently, however, immunoprecipitation experiments with antibodies to purified, rotenone-sensitive NADH-ubiquinone oxido-reductase [hereafter referred to as respiratory chain NADH dehydrogenase or complex I] from bovine heart, as well as enzyme fractionation studies, have indicated that six human URF's (that is, URF1, URF2, URF3, URF4, URF4L, and URF5, hereafter referred to as ND1, ND2, ND3, ND4, ND4L, and ND5) encode subunits of complex I. This is a large complex that also contains many subunits synthesized in the cytoplasm.*

[*The full paragraph includes one more sentence: "Support for such functional identification of the URF products has come from the finding that the purified rotenone-sensitive NADH dehydrogenase from *Neurospora crassa* contains several subunits synthesized within the mitochondria, and from the observation that the stopper mutant of *Neurospora crassa*, whose mtDNA lacks two genes homologous to URF2 and URF3, has no functional complex I." We have omitted this sentence both because the passage is long enough as is and because it raises no additional structural issues.]

Ask any ten people why this paragraph is hard to read, and nine are sure to mention the technical vocabulary; several will also suggest that it requires specialized background knowledge. Those problems turn out to be only a small part of the difficulty. Here is the passage again, with the difficult words temporarily lifted:

> The smallest of the URF's, and [A], has been identified as a [B] subunit 8 gene. The functional significance of the other URF's has been, on the contrary, elusive. Recently, however, [C] experiments, as well as [D] studies, have indicated that six human URF's [1-6] encode subunits of Complex I. This is a large complex that also contains many subunits synthesized in the cytoplasm.

It may now be easier to survive the journey through the prose, but the passage is still difficult. Any number of questions present themselves: What has the first sentence of the passage to do with the last sentence? Does the third sentence contradict what we have been told in the second sentence? Is the functional significance of URF's still "elusive"? Will this passage lead us to further discussion about URF's, or about Complex I, or both?

Information is interpreted more easily and more uniformly if it is placed where most readers expect to find it.

Knowing a little about the subject matter does not clear up all the confusion. The intended audience of this passage would probably possess at least two items of essential technical information: first, "URF" stands for "Uninterrupted Reading Frame," which describes a segment of DNA organized in such a way that it could encode a protein, although no such protein product has yet been identified; second, both APTase and NADH oxido-reductase are enzyme complexes central to energy metabolism. Although this information may provide some sense of comfort, it does little to answer the interpretive questions that need answering. It seems the reader is hindered by more than just the scientific jargon.

To get at the problem, we need to articulate something about how readers go about reading. We proceed to the first of several reader expectations.

Subject-Verb Separation

Look again at the first sentence of the passage cited above. It is relatively long, 42 words; but that turns out not to be the main cause of its burdensome complexity. Long sentences need not be difficult to read; they are only difficult to write. We have seen sentences of over 100 words that flow easily and persuasively toward their clearly demarcated destination. Those well-wrought serpents all had something in common: Their structure presented information to readers in the order the readers needed and expected it.

Beginning with the exciting material and ending with a lack of luster often leaves us disappointed and destroys our sense of momentum.

The first sentence of our example passage does just the opposite: it burdens and obstructs the reader, because of an all-too-common structural defect. Note that the grammatical subject ("the smallest") is separated from its verb ("has been identified") by 23 words, more than half the sentence. Readers expect a grammatical subject to be followed immediately by the verb. Anything of length that intervenes between subject and verb is read as an interruption, and therefore as something of lesser importance.

The reader's expectation stems from a pressing need for syntactic resolution, fulfilled only by the arrival of the verb. Without the verb, we do not know what the subject is doing, or what the sentence is all about. As a result, the reader focuses attention on the arrival of the verb and resists recognizing anything in the interrupting material as being of primary importance. The longer the interruption lasts, the more likely it becomes that the "interruptive" material actually contains important information; but its structural location will continue to brand it as merely interruptive. Unfortunately, the reader will not discover its true value until too late—until the sentence has ended without having produced anything of much value outside of that subject-verb interruption.

In this first sentence of the paragraph, the relative importance of the intervening material is difficult to evaluate. The material might conceivably be quite significant, in which case the writer should have positioned it to reveal that importance. Here is one way to incorporate it into the sentence structure:

> The smallest of the URF's is URFA6L, a 207-nucleotide (nt) reading frame overlapping out of phase the NH_2-terminal portion of the
>
> adenosinetriphosphatase (ATPase) subunit 6 gene; it has been identified as the animal equivalent of the recently discovered yeast H^+-ATPase subunit 8 gene.

On the other hand, the intervening material might be a mere aside that diverts attention from more important ideas; in that case the writer should have deleted it, allowing the prose to drive more directly toward its significant point:

> The smallest of the URF's (URFA6L) has been identified as the animal equivalent of the recently discovered yeast H^+-ATPase subunit 8 gene.

Only the author could tell us which of these revisions more accurately reflects his intentions.

These revisions lead us to a second set of reader expectations. Each unit of discourse, no matter what the size, is expected to serve a single function, to make a single point. In the case of a sentence, the point is expected to appear in a specific place reserved for emphasis.

The Stress Position

It is a linguistic commonplace that readers naturally emphasize the material that arrives at the end of a sentence. We refer to that location as a "stress position." If a writer is consciously aware of this tendency, she can arrange for the emphatic information to appear at the moment the reader is naturally exerting the greatest reading emphasis. As a result, the chances greatly increase that reader and writer will perceive the same material as being worthy of

primary emphasis. The very structure of the sentence thus helps persuade the reader of the relative values of the sentence's contents.

The inclination to direct more energy to that which arrives last in a sentence seems to correspond to the way we work at tasks through time. We tend to take something like a "mental breath" as we begin to read each new sentence, thereby summoning the tension with which we pay attention to the unfolding of the syntax. As we recognize that the sentence is drawing toward its conclusion, we begin to exhale that mental breath. The exhalation produces a sense of emphasis. Moreover, we delight in being rewarded at the end of a labor with something that makes the ongoing effort worthwhile. Beginning with the exciting material and ending with a lack of luster often leaves us disappointed and destroys our sense of momentum. We do not start with the strawberry shortcake and work our way up to the broccoli.

When the writer puts the emphatic material of a sentence in any place other than the stress position, one of two things can happen; both are bad. First, the reader might find the stress position occupied by material that clearly is not worthy of emphasis. In this case, the reader must discern, without any additional structural clue, what else in the sentence may be the most likely candidate for emphasis. There are no secondary structural indications to fall back upon. In sentences that are long, dense or sophisticated, chances soar that the reader will not interpret the prose precisely as the writer intended. The second possibility is even worse: The reader may find the stress position occupied by something that does appear capable of receiving emphasis, even though the writer did not intend to give it any stress. In that case, the reader is highly likely to emphasize this imposter material, and the writer will have lost an important opportunity to influence the reader's interpretive process.

The stress position can change in size from sentence to sentence. Sometimes it consists of a single word; sometimes it extends to several lines. The definitive factor is this: The stress position coincides with the moment of syntactic closure. A reader has reached the beginning of the stress position when she knows there is nothing left in the clause or sentence but the material presently being read. Thus a whole list, numbered and indented, can occupy the stress position of a sentence if it has been clearly announced as being all that remains of that sentence. Each member of that list, in turn, may have its own internal stress position, since each member may produce its own syntactic closure.

Within a sentence, secondary stress positions can be formed by the appearance of a properly used colon or semicolon; by grammatical convention, the material preceding these punctuation marks must be able to stand by itself as a complete sentence. Thus, sentences can be extended effortlessly to dozens of words, as long as there is a medial syntactic closure for every piece of new, stress-worthy information along the way. One of our revisions of the initial sentence can serve as an example:

> The smallest of the URF's is URFA6L, a 207-nucleotide (nt) reading frame overlapping out of phase the NH_2-terminal portion of the adenosinetriphosphatase (ATPase) subunit 6 gene; it has been identified as the animal equivalent of the recently discovered yeast H^+-ATPase subunit 8 gene.

By using a semicolon, we created a second stress position to accommodate a second piece of information that seemed to require emphasis.

We now have three rhetorical principles based on reader expectations: First, grammatical subjects should be followed as soon as possible by their verbs; second, every unit of discourse, no matter the size, should serve a single function or make a single point; and, third, information intended to be emphasized should appear at points of syntactic closure. Using these principles, we can begin to unravel the problems of our example prose.

Note the subject-verb separation in the 62-word third sentence of the original passage:

> Recently, however, immunoprecipitation experiments with antibodies to purified, rotenone-sensitive NADH-ubiquinone oxido-reductase [hereafter referred to as respiratory chain NADH dehydrogenase or complex I] from bovine heart, as well as enzyme fractionation studies, have indicated that six human URF's (that is, URF1, URF2, URF3, URF4, URF4L, and URF5, hereafter referred to as ND1, ND2, ND3, ND4, ND4L and ND5) encode subunits of complex I.

After encountering the subject ("experiments"), the reader must wade through 27 words (including three hyphenated compound words, a parenthetical interruption and an "as well as" phrase) before alighting on the highly uninformative and disappointingly anticlimactic verb ("have indicated"). Without a moment to recover, the reader is handed a "that" clause in which the new subject ("six human URF's") is separated from its verb ("encode") by yet another 20 words.

If we applied the three principles we have developed to the rest of the sentences of the example, we could generate a great many revised versions of each. These revisions might differ significantly from one another in the way their structures indicate to the reader the various weights and balances to be given to the information. Had the author placed all stress-worthy material in stress positions, we as a reading community would have been far more likely to interpret these sentences uniformly.

We couch this discussion in terms of "likelihood" because we believe that meaning is not inherent in discourse by itself; "meaning" requires the combined participation of text and reader. All sentences are infinitely interpretable, given an infinite number of interpreters. As communities of readers, however, we tend to work out tacit agreements as to what kinds of meaning are most likely to be extracted from certain articulations. We cannot succeed in making even a single sentence mean one and only one thing; we can only increase the odds that a large majority of readers will tend to interpret our discourse according to our intentions. Such success will follow from authors becoming more consciously aware of the various reader expectations presented here.

We cannot succeed in making even a single sentence mean one and only one thing; we can only increase the odds that a large majority of readers will tend to interpret our discourse according to our intentions.

Here is one set of revisionary decisions we made for the example:

> The smallest of the URF's, URFA6L, has been identified as the animal equivalent of the recently discovered yeast H^+-ATPase subunit 8 gene; but the functional significance of other URF's has been more elusive. Recently, however, several human URF's have been shown to encode subunits of rotenone-sensitive NADH-ubiquinone oxido-reductase. This is a large complex that also contains many subunits synthesized in the cytoplasm; it will be referred to hereafter as respiratory chain NADH dehydrogenase or complex I. Six subunits of Complex I were shown by enzyme fractionation studies and immunoprecipitation experiments to be encoded by six human URF's (URF1, URF2, URF3, URF4, URF4L, and URF5); these URF's will be referred to subsequently as ND1, ND2, ND3, ND4, ND4L and ND5.

Sheer length was neither the problem nor the solution. The revised version is not noticeably shorter than the original; nevertheless, it is significantly easier to interpret. We have indeed deleted certain words, but not on the basis of wordiness or excess length. (See especially the last sentence of our revision.)

When is a sentence too long? The creators of readability formulas would have us believe there exists some fixed number of words (the favorite is 29) past which a sentence is too hard to read. We disagree. We have seen 10-word sentences that are virtually impenetrable and, as we mentioned above, 100-word sentences that flow effortlessly to their points of resolution. In place of the word-limit concept, we offer the following definition: A sentence is too long when it has more viable candidates for stress positions than there are stress positions available. Without the stress position's locational clue that its material is intended to be emphasized, readers are left too much to their own devices in deciding just what else in a sentence might be considered important.

In revising the example passage, we made certain decisions about what to omit and what to emphasize. We put subjects and verbs together to lessen the reader's syntactic burdens; we put the material we believed worthy of emphasis in stress positions; and we discarded material for which we could not discern significant connections. In doing so, we have produced a clearer passage—but not one that necessarily reflects the author's intentions; it reflects only our interpretation of the author's intentions. The more problematic the structure, the less likely it becomes that a grand majority of readers will perceive the discourse in exactly the way the author intended.

The information that begins a sentence establishes for the reader a perspective for viewing the sentence as a unit.

It is probable that many of our readers--and perhaps even the authors—will disagree with some of our choices. If so, that disagreement underscores our point: The original failed to communicate its ideas and their connections clearly. If we happened to have interpreted the passage as you did, then we can make a different point: No one should have to work as hard as we did to unearth the content of a single passage of this length.

The Topic Position

To summarize the principles connected with the stress position, we have the proverbial wisdom, "Save the best for last." To summarize the principles connected with the other end of the sentence, which we will call the topic position, we have its proverbial contradiction, "First things first." In the stress position the reader needs and expects closure and fulfillment; in the topic position the reader needs and expects perspective and context. With so much of reading comprehension affected by what shows up in the topic position, it behooves a writer to control what appears at the beginning of sentences with great care.

The information that begins a sentence establishes for the reader a perspective for viewing the sentence as a unit: Readers expect a unit of discourse to be a story about whoever shows up first. "Bees disperse pollen" and "Pollen is dispersed by bees" are two different but equally respectable sentences about the same facts. The first tells us something about bees; the second tells us something about pollen. The passivity of the second sentence does not by itself impair its quality; in fact, "Pollen is dispersed by bees" is the superior sentence if it appears in a paragraph that intends to tell us a continuing story about pollen. Pollen's story at that moment is a passive one.

Readers also expect the material occupying the topic position to provide them with linkage (looking backward) and context (looking forward). The information in the topic position prepares the reader for upcoming material by connecting it backward to the previous discussion. Although linkage and context can derive from several sources, they stem primarily from material that the reader has already encountered within this particular piece of discourse. We refer to this familiar, previously introduced material as "old information." Conversely, material making its first appearance in a discourse is "new information." When new information is important enough to receive emphasis, it functions best in the stress position.

When old information consistently arrives in the topic position, it helps readers to construct the logical flow of the argument: It focuses attention on one particular strand of the discussion, both harkening backward and leaning forward. In contrast, if the topic position is constantly occupied by material that fails to establish linkage and context, readers will have difficulty perceiving both the connection to the previous sentence and the projected role of the new sentence in the development of the paragraph as a whole.

Here is a second example of scientific prose that we shall attempt to improve in subsequent discussion:

> Large earthquakes along a given fault segment do not occur at random intervals because it takes time to accumulate the strain energy for the rupture. The rates at which tectonic plates move and accumulate strain at their boundaries are approximately uniform. Therefore, in first approximation, one may expect that large ruptures of the same fault segment will occur at approximately constant time intervals. If subsequent main shocks have different amounts of slip across the fault, then the recurrence time may vary, and the basic idea of periodic mainshocks must be modified. For great plate boundary ruptures the length and slip often vary by a factor of 2. Along the southern segment of the San Andreas fault the recurrence interval is 145 years with variations of several decades. The smaller the standard deviation of the average recurrence interval, the more specific could be the long term prediction of a future mainshock.

This is the kind of passage that in subtle ways can make readers feel badly about themselves. The individual sentences give the impression of being intelligently fashioned: They are not especially long or convoluted; their vocabulary is appropriately professional but not beyond the ken of educated general readers; and they are free of grammatical and dictional errors. On first reading, however, many of us arrive at the paragraph's end without a clear sense of where we have been or where we are going. When that happens, we tend to berate ourselves for not having paid close enough attention. In reality, the fault lies not with us, but with the author.

We can distill the problem by looking closely at the information in each sentence's topic position:

> Large earthquakes
> The rates
> Therefore...one
> subsequent mainshocks
> great plate boundary ruptures
> the southern segment of the San Andreas fault
> the smaller the standard deviation...

Much of this information is making its first appearance in this paragraph—in precisely the spot where the reader looks for old, familiar information. As a result, the focus of the story constantly shifts. Given just the material in the topic positions, no two readers would be likely to construct exactly the same story for the paragraph as a whole.

If we try to piece together the relationship of each sentence to its neighbors, we notice that certain bits of old information keep reappearing. We hear a good deal about the recurrence time between earthquakes: The first sentence introduces the concept of nonrandom intervals between earthquakes; the second sentence tells us that recurrence rates due to the movement of tectonic plates are more or less uniform; the third sentence adds that the recurrence rates of major earthquakes should also be somewhat predictable; the fourth sentence adds that recurrence rates vary with some conditions; the fifth sentence adds information about one particular variation; the sixth sentence adds a recurrence-rate example from California; and the last sentence tells us something about how recurrence rates can be described statistically. This refrain of "recurrence intervals" constitutes the major string of old information in the paragraph. Unfortunately, it rarely appears at the beginning of sentences, where it would help us maintain our focus on its continuing story.

In reading, as in most experiences, we appreciate the opportunity to become familiar with a new environment before having to function in it. Writing that continually begins sentences with new information and ends with old information forbids both the sense of comfort and orientation at the start and the sense of fulfilling arrival at the end. It misleads the reader as to whose story is being told; it burdens the reader with new information that must be carried further into the sentence before it can be connected to the discussion; and it creates ambiguity as to which material the writer intended the reader to emphasize. All of these distractions require that readers expend a disproportionate amount of energy to unravel the structure of the prose, leaving less energy available for perceiving content.

We can begin to revise the example by ensuring the following for each sentence:

1. The backward-linking old information appears in the topic position.
2. The person, thing or concept whose story it is appears in the topic position.
3. The new, emphasis-worthy information appears in the stress position.

Once again, if our decisions concerning the relative values of specific information differ from yours, we can all blame the author, who failed to make his intentions apparent. Here first is a list of what we perceived to be the new, emphatic material in each sentence:

> time to accumulate strain energy along a fault
> approximately uniform
> large ruptures of the same fault
> different amounts of slip
> vary by a factor of 2
> variations of several decades

predictions of future mainshock

Now, based on these assumptions about what deserves stress, here is our proposed revision:

> Large earthquakes along a given fault segment do not occur at random intervals because it takes time to accumulate the strain energy for the rupture. The rates at which tectonic plates move and accumulate strain at their boundaries are roughly uniform. Therefore, nearly constant time intervals (at first approximation) would be expected between large ruptures of the same fault segment. [However?], the recurrence time may vary; the basic idea of periodic mainshocks may need to be modified if subsequent mainshocks have different amounts of slip across the fault. [Indeed?], the length and slip of great plate boundary ruptures often vary by a factor of 2. [For example?], the recurrence intervals along the southern segment of the San Andreas fault is 145 years with variations of several decades. The smaller the standard deviation of the average recurrence interval, the more specific could be the long term prediction of a future mainshock.

Many problems that had existed in the original have now surfaced for the first time. Is the reason earthquakes do not occur at random intervals stated in the first sentence or in the second? Are the suggested choices of "however," "indeed," and "for example" the right ones to express the connections at those points? (All these connections were left unarticulated in the original paragraph.) If "for example" is an inaccurate transitional phrase, then exactly how does the San Andreas fault example connect to ruptures that "vary by a factor of 2"? Is the author arguing that recurrence rates must vary because fault movements often vary? Or is the author preparing us for a discussion of how in spite of such variance we might still be able to predict earthquakes? This last question remains unanswered because the final sentence leaves behind earthquakes that recur at variable intervals and switches instead to earthquakes that recur regularly. Given that this is the first paragraph of the article, which type of earthquake will the article most likely proceed to discuss? In sum, we are now aware of how much the paragraph had not communicated to us on first reading. We can see that most of our difficulty was owing not to any deficiency in our reading skills but rather to the author's lack of comprehension of our structural needs as readers.

In our experience, the misplacement of old and new information turns out to be the No. 1 problem in American professional writing today.

In our experience, the misplacement of old and new information turns out to be the No. 1 problem in American professional writing today. The source of the problem is not hard to discover: Most writers produce prose linearly (from left to right) and through time. As they begin to formulate a sentence, often their primary anxiety is to capture the important new thought before it escapes. Quite naturally they rush to record that new information on paper, after which they can produce at their leisure contextualizing material that links back to the previous discourse. Writers who do this consistently are attending more to their own need for unburdening themselves of their information than to the reader's need for receiving the material. The methodology of reader expectations articulates the reader's needs explicitly, thereby making writers consciously aware of structural problems and ways to solve them.

Put in the topic position the old information that links backward; put in the stress position the new information you want the reader to emphasize.

A note of clarification: Many people hearing this structural advice tend to oversimplify it to the following rule: "Put the old information in the topic position and the new information in the stress position." No such rule is possible. Since by definition all information is either old or new, the space between the topic position and the stress position must also be filled with old and new information. Therefore the principle (not rule) should be stated as follows: "Put in the topic position the old information that links backward; put in the stress position the new information you want the reader to emphasize."

Perceiving Logical Gaps

When old information does not appear at all in a sentence, whether in the topic position or elsewhere, readers are left to construct the logical linkage by themselves. Often this happens when the connections are so clear in the writer's mind that they seem unnecessary to state; at those moments, writers underestimate the difficulties and ambiguities inherent in the reading process. Our third example attempts to demonstrate how paying attention to the placement of old and new information can reveal where a writer has neglected to articulate essential connections.

> The enthalpy of hydrogen bond formation between the nucleoside bases 2'deoxyguanosine (dG) and 2'deoxycytidine (dC) has been determined by direct measurement. dG and dC were derivatized at the 5' and 3' hydroxyls with triisopropylsilyl groups to obtain solubility of the nucleosides in non-aqueous solvents and to prevent the ribose hydroxyls from forming hydrogen bonds. From isoperibolic titration measurements, the enthalpy of dC:dG base pair formation is -6.65±0.32 kcal/mol.

Although part of the difficulty of reading this passage may stem from its abundance of specialized technical terms, a great deal more of the difficulty can be attributed to its structural problems. These problems are now familiar: We are not sure at all times whose story is being told; in the first sentence the subject and verb are widely separated; the second sentence has only one stress position but two or three pieces of information that are probably worthy of emphasis —"solubility ...solvents," "prevent... from forming hydrogen bonds" and perhaps "triisopropylsilyl groups." These perceptions suggest the following revision tactics:

> 1. Invert the first sentence, so that (a) the subject-verb-complement connection is unbroken, and (b) "dG" and "dC" are introduced in the stress position as new and interesting information. (Note that inverting the sentence requires stating who made the measurement; since the authors performed the first direct measurement, recognizing their agency in the topic position may well be appropriate.)
>
> 2. Since "dG and "dC" become the old information in the second sentence, keep them up front in the topic position.
>
> 3. Since "triisopropylsilyl groups" is new and important information here, create for it a stress position.
>
> 4. "Triisopropylsilyl groups" then becomes the old information of the clause in which its effects are described; place it in the topic position of this clause.
>
> 5. Alert the reader to expect the arrival of two distinct effects by using the flag word "both." "Both" notifies the reader that two pieces of new information will arrive in a single stress position.

Here is a partial revision based on these decisions:

> We have directly measured the enthalpy of hydrogen bond formation between the nucleoside bases 2'deoxyguanosine (dG) and 2'deoxycytidine (dC). dG and dC were derivatized at the 5' and 3' hydroxyls with triisopropylsilyl groups; these groups serve both to solubilize the nucleosides in non-aqueous solvents and to prevent the ribose hydroxyls from forming hydrogen bonds. From isoperibolic titration measurements, the enthalpy of dC:dG base pair formation is -6.65±0.32 kcal/mol.

The outlines of the experiment are now becoming visible, but there is still a major logical gap. After reading the second sentence, we expect to hear more about the two effects that were important enough to merit placement in its stress position. Our expectations are frustrated, however, when those effects are not mentioned in the next sentence: "From isoperibolic titration measurements, the enthalpy of dC:dG base pair formation is -6.65±0.32 kcal/mol." The authors have neglected to explain the relationship between the derivatization they performed (in the second sentence) and the measurements they made (in the third sentence). Ironically, that is the point they most wished to make here.

At this juncture, particularly astute readers who are chemists might draw upon their specialized knowledge, silently supplying the missing connection. Other readers are left in the dark. Here is one version of what we think the authors meant to say, with two additional sentences supplied from a knowledge of nucleic acid chemistry:

> We have directly measured the enthalpy of hydrogen bond formation between the nucleoside bases 2'deoxyguanosine (dG) and 2'deoxycytidine (dC). dG and dC were derivatized at the 5' and 3' hydroxyls with triisopropylsiyl groups; these groups serve both to solubilize the nucleosides in non-aqueous solvents and to prevent the ribose hydroxyls from forming hydrogen bonds. Consequently, when the derivatized nucleosides are dissolved in non-aqueous solvents, hydrogen bonds form almost exclusively between the bases. Since the interbase hydrogen bonds are the only bonds to form upon mixing, their

> enthalpy of formation can be determined directly by measuring the enthalpy of mixing. From our isoperibolic titration measurements, the enthalpy of dG:dC base pair formation is -6.65±0.32 kcal/mol.

Each sentence now proceeds logically from its predecessor. We never have to wander too far into a sentence without being told where we are and what former strands of discourse are being continued. And the "measurements" of the last sentence has now become old information, reaching back to the "measured directly" of the preceding sentence. (It also fulfills the promise of the "we have directly measured" with which the paragraph began.) By following our knowledge of reader expectations, we have been able to spot discontinuities, to suggest strategies for bridging gaps, and to rearrange the structure of the prose, thereby increasing the accessibility of the scientific content.

Locating the Action

Our final example adds another major reader expectation to the list.

> Transcription of the 5*S* RNA genes in the egg extract is TFIIIA-dependent. This is surprising, because the concentration of TFIIIA is the same as in the oocyte nuclear extract. The other transcription factors and RNA polymerase III are presumed to be in excess over available TFIIIA, because tRNA genes are transcribed in the egg extract. The addition of egg extract to the oocyte nuclear extract has two effects on transcription efficiency. First, there is a general inhibition of transcription that can be alleviated in part by supplementation with high concentrations of RNA polymerase III. Second, egg extract destabilizes transcription complexes formed with oocyte but not somatic 5*S* RNA genes.

The barriers to comprehension in this passage are so many that it may appear difficult to know where to start revising. Fortunately, it does not matter where we start, since attending to any one structural problem eventually leads us to all the others.

We can spot one source of difficulty by looking at the topic positions of the sentences: We cannot tell whose story the passage is. The story's focus (that is, the occupant of the topic position) changes in every sentence. If we search for repeated old information in hope of settling on a good candidate for several of the topic positions, we find all too much of it: egg extract, TFIIIA, oocyte extract, RNA polymerase III, 5*S* RNA, and transcription. All of these reappear at various points, but none announces itself clearly as our primary focus. It appears that the passage is trying to tell several stories simultaneously, allowing none to dominate.

We are unable to decide among these stories because the author has not told us what to do with all this information. We know who the players are, but we are ignorant of the actions they are presumed to perform. This violates yet another important reader expectation: Readers expect the action of a sentence to be articulated by the verb.

Here is a list of the verbs in the example paragraph:

> is
> is...is
> are presumed to be
> are transcribed
> has
> is...can be alleviated
> destabilizes

The list gives us too few clues as to what actions actually take place in the passage. If the actions are not to be found in the verbs, then we as readers have no secondary structural clues for where to locate them. Each of us has to make a personal interpretive guess; the writer no longer controls the reader's interpretive act.

As critical scientific readers, we would like to concentrate our energy on whether the experiments prove the hypotheses.

Worse still, in this passage the important actions never appear. Based on our best understanding of this material, the verbs that connect these players are "limit" and "inhibit." If we express those actions as verbs and place the most frequently occurring information—"egg extract" and "TFIIIA"—in the topic position whenever possible,* we can generate the following revision:

> In the egg extract, the availability of TFIIIA limits transcription of the 5*S* RNA genes. This is surprising because the same concentration of TFIIIA does not limit transcription in the oocyte nuclear extract. In the egg extract, transcription is not limited by RNA polymerase or other factors because transcription of tRNA genes indicates that these factors are in excess over available TFIIIA. When added to the nuclear extract, the egg extract affected the efficiency of transcription in two ways. First, it inhibited transcription generally; this inhibition could be alleviated in part by supplementing the mixture with high concentrations of RNA polymerase III. Second, the egg extract destabilized transcription complexes formed by oocyte but not by somatic 5*S* genes.

[*We have chosen these two pieces of old information as the controlling contexts for the passage. That choice was neither arbitrary nor born of logical necessity; it was simply an act of interpretation. All readers make exactly that kind of choice in the reading of every sentence. The fewer the structural clues to interpretation given by the author, the more variable the resulting interpretations will tend to be.]

As a story about "egg extract," this passage still leaves something to be desired. But at least now we can recognize that the author has not explained the connection between "limit" and "inhibit." This unarticulated connection seems to us to contain both of her hypotheses: First, that the limitation on transcription is caused by an inhibitor of TFIIIA present in the egg extract; and, second, that the action of that inhibitor can be detected by adding the egg extract to the oocyte extract and examining the effects on transcription. As critical scientific readers, we would like to concentrate our energy on whether the experiments prove the hypotheses. We cannot begin to do so if we are left in doubt as to what those hypotheses might be—and if we are using most of our energy to discern the structure of the prose rather than its substance.

Writing and the Scientific Process

We began this article by arguing that complex thoughts expressed in impenetrable prose can be rendered accessible and clear without minimizing any of their complexity. Our examples of scientific writing have ranged from the merely cloudy to the virtually opaque; yet all of them could be made significantly more comprehensible by observing the following structural principles:

1. Follow a grammatical subject as soon as possible with its verb.
2. Place in the stress position the "new information" you want the reader to emphasize.
3. Place the person or thing whose "story" a sentence is telling at the beginning of the sentence, in the topic position.
4. Place appropriate "old information" (material already stated in the discourse) in the topic position for linkage backward and contextualization forward.
5. Articulate the action of every clause or sentence in its verb.
6. In general, provide context for your reader before asking that reader to consider anything new.
7. In general, try to ensure that the relative emphases of the substance coincide with the relative expectations for emphasis raised by the structure.

It may seem obvious that a scientific document is incomplete without the interpretation of the writer; it may not be so obvious that the document cannot "exist" without the interpretation of each reader.

None of these reader-expectation principles should be considered "rules." Slavish adherence to them will succeed no better than has slavish adherence to avoiding split infinitives or to using the active voice instead of the passive. There can be no fixed algorithm for good writing, for two reasons. First, too many reader expectations are functioning at any given moment for structural decisions to remain clear and easily activated. Second, any reader expectation can be violated to good effect. Our best stylists turn out to be our most skillful violators; but in order to carry this off, they must fulfill expectations most of the time, causing the violations to be perceived as exceptional moments, worthy of note.

A writer's personal style is the sum of all the structural choices that person tends to make when facing the challenges of creating discourse. Writers who fail to put new information in the stress position of many sentences in one document are likely to repeat that unhelpful structural pattern in all other documents. But for the very reason that writers tend to be consistent in making such choices, they can learn to improve their writing style; they can permanently reverse those habitual structural decisions that mislead or burden readers.

We have argued that the substance of thought and the expression of thought are so inextricably intertwined that changes in either will affect the quality of the other. Note that only the first of our examples (the paragraph about URF's) could be revised on the basis of the methodology to reveal a nearly finished passage. In all the other examples, revision revealed existing conceptual gaps and other problems that had been submerged in the originals by dysfunctional structures. Filling the gaps required the addition of extra material. In revising each of these examples, we arrived at a point where we could proceed no further without either supplying connections between ideas or eliminating some existing material altogether. (Writers who use reader-expectation principles on their own prose will not have to conjecture or infer; they know what the prose is intended to convey.) Having begun by analyzing the structure of the prose, we were led eventually to reinvestigate the substance of the science.

The substance of science comprises more than the discovery and recording of data; it extends crucially to include the act of interpretation. It may seem obvious that a scientific document is incomplete without the interpretation of the writer; it may not be so obvious that the document cannot "exist" without the interpretation of each reader. In other words, writers cannot "merely" record data, even if they try. In any recording or articulation, no matter how haphazard or confused, each word resides in one or more distinct structural locations. The resulting structure, even more than the meanings of individual words, significantly influences the reader during the act of interpretation. The question then becomes whether the structure created by the writer (intentionally or not) helps or hinders the reader in the process of interpreting the scientific writing.

The writing principles we have suggested here make conscious for the writer some of the interpretive clues readers derive from structures. Armed with this awareness, the writer can achieve far greater control (although never complete control) of the reader's interpretive process. As a concomitant function, the principles simultaneously offer the writer a fresh re-entry to the thought process that produced the science. In real and important ways, the structure of the prose becomes the structure of the scientific argument. Improving either one will improve the other.

The methodology described in this article originated in the linguistic work of Joseph M. Williams of the University of Chicago,Gregory G. Colomb of the Georgia Institute of Technology and George D. Gopen. Some of the materials presented here were discussed and developed in faculty writing workshops held at the Duke University Medical School.

Bibliography

- Colomb, Gregory G., and Joseph M. Williams. 1985. Perceiving structure in professional prose: a multiply determined experience. In *Writing in Non-Academic Settings,* eds. Lee Odell and Dixie Goswami. Guilford Press, pp. 87-128.
- Gopen, George D. 1987. Let the buyer in ordinary course of business beware: suggestions for revising the language of the Uniform Commercial Code. *University of Chicago Law Review* 54:1178-1214.
- Gopen, George D. 1990. *The Common Sense of Writing: Teaching Writing from the Reader's Perspective*.
- Williams, Joseph M. 1988. *Style: Ten Lessons in Clarity and Grace.* Scott, Foresman, & Co.