required to build the representation. Characterizations of coherence may be useful in controlling the analysis, and considering the form of reasoning underlying the discourse may help to characterize the form. As the book focuses on discourse, particular issues such as reference resolution and the maxims of conversation are highlighted. There are in fact particular questions addressed in some of the chapters which are especially relevant to certain computational linguistics research efforts.

The first three chapters present conceptions of the form of reasoning underlying discourse, especially arguments. These papers are relevant to computational linguists involved in constructing models for the analysis of discourse. Johnson-Laird argues that logical form has no role in accounting for deductive competence. Connectives and quantifiers do not merit a special treatment; people need only know the truth conditions of these terms in order to make deductions. In another chapter, Moore presents some evidence for induction as a model of reasoning. In contrast, Allwood claims that speakers share a normative intuition, following traditional logic, of the shape of an argument. He contributes some insights as well into the role of utterance-level intentions in discourse structure. Hagert and Waern present some insight into the form of invalid plans underlying discourse. They comment on the need to distinguish inferences underlying actual sentences from those used for inferencing (i.e., the point of view of the observer and the speaker).

The last three chapters of the book, by Kempson, Wilson and Sperber, and Wilks, present an interesting discussion of the procedures for discourse processing. Kempson draws on some suggestions of Wilson and Sperber to discuss the relationship between semantic and pragmatic processing, with an application to anaphora resolution. She proposes a mapping from surface structure to a logical form, which then interacts with a pragmatic, relevance-driven rule of antecedent identification. Wilson and Sperber address the use of "relevance" to utterance-level analysis, which is seen as a process of hypothesis formation. Wilks then criticizes Sperber and Wilson for failing to distinguish beliefs of conversants, in interpreting inference in discourse.

In general, the collection is a useful reference. My main negative comment is that some of the papers do not include enough examples (which would be eminently useful for people constructing implementations), and as a result end up appearing too general, too superficial. But on the whole I continue to have faith that dialog between cognitive scientists (psychologists, linguists) and computer science researchers is possible, even for those computer science people without the aim of having a cognitively accurate model of human processing. The class of input to process can be made clearer, and intuitions for the characterizations of processing models can be provided.

Robin Cohen's research concentrates on the structure of argument and discourse. Cohen's address is: Department of Computer Science, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1. E-mail: rcohen@watdragon.waterloo.edu

---

Language and Information

Zellig Sabbettai Harris
(University of Pennsylvania)

(Bampton lectures in America 28)
New York: Columbia University Press, 1988, ix+120 pp.
ISBN 0-231-06662-7; $20.00 (hb)

Reviewed by
Bruce Nevin
BBN Communications Corporation

The glib freedom with which we use the word *information* would lead one to suppose we know what we are talking about. Alas, not so. In a field that concerns itself with "information processing", it is remarkable if not embarrassing that there is still, after 40 years, no generally accepted, coherent definition of information to underwrite the enterprise.

It is well known that information theory is not concerned with the information content or meanings of particular texts or utterances. It interprets certain measures of probability or uncertainty in an ensemble of signal sequences (which may indeed be meaningless) as a metric of the difficulty of transmitting a given signal sequence, and then calls this metric, in a notoriously misleading way, the "amount of information" in the signal.[1]

Carnap and Bar-Hillel[2] announced long ago what was essentially a ramification of Carnap's work in inductive logic and probability, a **Theory of Semantic Information** dealing solely with linguistic entities ("state descriptions" in some logical language) and what they stand for or designate. Carnap's aim was to devise measures of "semantic content" that would enable him to get at "confirmation functions" to underwrite inductive logic. Bar-Hillel's initial enthusiasm was to develop a perhaps broader "calculus of information." Although the banner they dropped was taken up in the '60s by Hintikka and others,[3] it is safe to say that this line of thought has contributed little to a satisfactory definition of information.

Today, we witness the spectacle of Dretske and the situation semantics folks[4] mounted precariously on the Scylla of naive realism, tilting with Fodor atop the Charybdis of a mental representationalism that is philosophically more sophisticated but no less ad hoc in its misuse of metaphor.[5] Unfortunately, a summary of the well-deserved doubt that each casts upon the merits of the other's case is beyond the scope of this review.

The present book is a brief and very clear introduction to a body of work[6] that threads a naturalist path between these extremes and offers real insight into the nature of linguistic information. What is meant by this is the literal, objective information in discourse, as distinct from, e.g., gestural systems such as expressive intonation and other body language. The paradigmatic case is a technical paper in a subfield of science, as opposed to artistic expressions, such as music, dance, and literary and poetic uses of language.

Harris shows how natural language differs in many important respects not only from such gestural systems on the one hand, but also from mathematics, logic, and formal languages, on the other. The formal structure of operators and arguments that Harris finds in language resembles functors of a categorial grammar in logic, but contrasts with them in a number of ways, including the following:

- Words are classified as operators and arguments not ostensively, by listing them, but with respect to the argument requirement of their arguments. This dependency on dependency, by virtue of which we recognize language to be a mathematical object (p. 89), makes possible the striking simplicity of the theory.
- Arguments of an operator may occur in various orders relative to each other (such as topicalization, fronting of a non-first argument), though in most languages there are one or two preferred sequences.
- An operator may enter at various points relative to its arguments as a sentence is constructed, though again the normal choices are limited in most languages.
- These alternative linearizations include interruption—under a paratactic conjunction whose subordinating intonation is represented by a semicolon or, as here, by dashes—of a sentence by a secondary sentence. This is the origin in many languages of the relative clause and thence of all modifiers.
- Particular word combinations (operator-argument cooccurrences) are graded as to likelihood.
- Word occurrences that contribute little information (because of high likelihood) may occur in reduced phonemic form, even zero; conversely, occurrences with especially low likelihood may block otherwise customary reductions.

To see why approaches to information from the point of view of mathematical logic have been unable to get at intuitively appealing notions of information based upon our everyday use of natural language, we must see just how and why language "carries" information.

How can a formal theory of syntax—formal in that it defines entities by their frequency of occurrence relative to each other rather than by their phonetic or semantic properties—have as its result (and indeed its point) an account of meaning? The answer lies in the well-known relation of information to redundancy or expectability. A central point is that *there is no external metalanguage* for the investigation of language, as there is for every other science. The information in language can be represented and explained only in language itself. All that is available for accomplishing this is to exploit the departures from randomness in language, first to distinguish its elements, and then to determine the structures (patterns of redundancy) in it. But it is precisely this redundancy among its elements that language itself uses for informational purposes: information is present in a text because the elements of language do not occur randomly with respect to each other.

For this reason, it is of critical importance that the description introduce no extrinsic redundancy: that it employ the fewest and simplest entities and the fewest and simplest rules, with (if possible) no repetition.

> The notorious complexity of grammar, most of which is created by the reductions, is not due to complexity in the information and is not needed for information (p.29).

> [A]s we approach a least grammar, with least redundancy in the description of the structure, the connection of that grammar with information becomes much stronger. Indeed, the step-by-step connection of information with structure is found to be so strong as to constitute a test of the relevance of any proposed structural analysis of language. . . . the components that go into the making of the structure are the components that go into the making of the information (p.57).[7]

Having arrived at a "least grammar", Harris shows us that a representation of the grammar of a text is also a representation of the information in it. For example, he shows us how analysis of texts in a science sublanguage yields what he calls a science language, "a body of canonical formulas, representing the science statements after synonymy and the paraphrastic reductions have been undone" (p.51). It is most striking that this representation of the information in technical articles is the same regardless of whether the original language was English, French, or some other language: its structure is a characteristic of the science and not of the particular natural language the investigators used for reporting their results and from which it was derived. Needless to say, this is a matter of some interest for machine translation, information retrieval, and knowledge representation.[8]

The significant redundancy in language has two sources, two constraints on the equiprobability of word combinations.

- The first constraint, which creates sentence structure, is the partial ordering of operators with respect to their arguments. Its significance is, roughly, predication: an operator is said "of" its arguments.
- The second constraint, which specifies word meanings, is on the relative likelihood of particular constructions of operator and argument words.[9] As noted above, an operator-argument pair (which are always adjacent at time of entry—a matter of some computational importance) may have exceptionally high

likelihood (low information), above average or "selectional" likelihood, lower likelihood, or exceptionally low likelihood.[10]

As described above, words entering in the ongoing construction of a sentence are given a particular linear order. If with newly entering words a high-likelihood collocation arises, a reduction may (usually optionally) produce a more compact alternant form of the sentence. The reductions constitute a third constraint on co-occurrences of word shapes (allomorphs), but not one that contributes to information, since it is precisely low-information word occurrences that are affected.[11] With both the alternant linearizations and the reductions, what changes is emphasis or ease of access to the objective information, which remains invariant. Nuances of meaning expressed by these means or by pauses, gesture, and so on, can also be expressed by using the above two "substantive" constraints in explicit albeit perhaps more awkward language, as anyone knows who has puzzled out a joke or an "untranslatable" idiom in a foreign language.

Every increment of information in a text corresponds to a step of sentence construction exercising one of these constraints. There is no a priori structure of information onto which grammar maps the spoken (or written) words of language: rather, the information in a text inheres in and is the natural interpretation of the structure of words that enables it to be expressed. *Referring* appears to be a matter of a loose correspondence between the redundancies in a text and similar departures from randomness in a set of events (pp. 83–85).

These are some of the chief themes of the first three lectures, entitled respectively "A Formal Theory of Syntax", "Science Sublanguages", and "Information". The fourth lecture, "The Nature of Language", is more far-ranging in content, discussing the structural properties of language, including language universals; language change and different aspects of language that are in greater or lesser degree subject to it; and stages and processes in the origin and development of language based upon the several contributory information-making constraints described earlier. In the final section on "Language as an Evolving System", Harris shows how language likely evolved and is evolving: "We may still be at an early stage of it" (p.107).

In the closing pages, Harris responds to rationalist claims that complex, species-specific, innate biological structures are necessary for something as complex as language to be learnable, arguing that

> there is nothing magical about how much, and what, is needed in order to speak . . . . We can see roughly what kind of mental capacity is involved in knowing each contribution to the structure . . . . The kind of knowing that is needed here is not as unique as language seems to be, and not as ungraspable in amount.
>
> The overall picture that we obtain is of a self-organizing system growing out of real-life conditions in combining sound sequences. Indeed, it could hardly be otherwise, since there is no external metalanguage in which to define the structure, and no external agent to have created it (pp. 112–113).

This book is a clear and succinct summation in compact form of an extensive body of scientific investigation that no one interested in either language or information can afford to ignore.[12]

References

Bar-Hillel, J. 1952 Semantic Information and its Measures. In von Foerster, H. (Ed.). *Cybernetics*; Transactions of the 8th Conference. Josiah Macy Foundation, New York, NY:33–48. Reprinted in Bar-Hillel (1964).

Bar-Hillel, J. 1964 *Language and Information. Selected Essays on Their Theory and Application.* Addison-Wesley, Reading, MA.

Carnap, R. and Bar-Hillel, Y. 1952 *An Outline of a Theory of Semantic Information.* Technical Report No. 247, Research Lab. of Electronics, MIT. Cambridge, MA. Reprinted in Bar-Hillel (1964).

Dretske, F. I. 1981 *Knowledge and the Flow of Information.* MIT Press, Cambridge, MA.

Dretske, F. I. 1983 Précis of *Knowledge and the Flow of Information. The Behavioral and Brain Sciences* 6:55–90.

Fodor, J. A. 1986 Information and Association. *Notre Dame Journal of Formal Logic* 27(3):307–323.

Fodor, J. A. (forthcoming). What is Information? Paper delivered to the American Philosophical Association.

Harris, Z. S. 1954 Distributional Structure. *WORD* 10:146–162. Reprinted in J. Fodor and J. Katz, *The Structure of Language: Readings in the Philosophy of Language*, Prentice-Hall, 1964. Reprinted in Z. S. Harris, *Papers in Structural and Transformational Linguistics*, Reidel, Dordrecht, 1970, 775–794.

Harris, Z. S. 1982 *A Grammar of English on Mathematical Principles.* Wiley/Interscience, New York, NY.

Harris, Z. S.; Gottfried, M.; Ryckman, T. et al (in press). *The Form of Information in Science.* D. Reidel, Dordrecht.

Hintikka, J. and Suppes, P. (eds.). 1970 *Information and Inference.* Humanities Press, New York, NY.

Israel, D. and Perry, J. (forthcoming). *What is Information?* Center for the Study of Language and Information, Stanford University, Stanford, CA.

Johnson, S. B. 1987 *An Analyzer for the Information Content of Sentences.* Ph.D. diss., New York University, New York, NY.

Ryckman, T. A. 1986 *Grammar and Information: An Investigation in Linguistic Metatheory.* Ph.D. diss., Columbia University, New York, NY.

Schützenberger, M-P. 1956 On Some Measures of Information Used in Statistics In Cherry, C. (ed.). *Information Theory*; Proceedings of a Symposium. Academic Press, New York, NY, 18–24.

Bruce Nevin is a manager and a researcher in interface design and information retrieval issues in the customer documentation department of BBN Communications Corporation. He studied linguistics at the University of Pennsylvania (where Zellig Harris was one of his lecturers), receiving the A.B. (1968) and A.M. (1970) degrees. He did further graduate work in linguistics at UC Berkeley in 1970–74. Nevin's address is BBN Communications Corporation, Mail Stop 045, 50 Moulton Street, Cambridge, MA 02238. E-mail: bn@cch.bbn.com.

Notes

1. For a comparison of the different measures of information in statistics and in communication theory—the more accurate

name—see Schützenberger (1956). For a summary of the issues, see Ryckman (1986), chap. 5.

2. Carnap and Bar-Hillel (1952), Bar-Hillel (1952). The present book seems in part responsive to this program, having the same title as Bar-Hillel (1964).
3. See papers collected in Hintikka and Suppes (1970).
4. Dretske (1981), Israel and Perry (forthcoming). Peer commentary in Dretske (1983), especially that of Haber, did not accept Dretske's attempted analogies to the metrics of Shannon and Weaver. The notion of "information pickup" implies a pre-established harmony of the world and the mind, disregarding the well-known arbitrariness of language.
5. While Fodor (1986) does gives a cogent criticism of attempts to locate information "in the world", the alternative "intentional" conception that he advocates relies on questionable assumptions of an "internal code" wherein such information is "encoded". The problem, of course, lies in unpacking this metaphor. Falling into the custom of taking the computational metaphor of mind literally, he resuscitates our old familiar homunculus (in computational disguise as the "executive") to provide a way out of the problem of node labels being of higher logical type than the nodes that they label. A simpler resolution follows from Harris's recognition that natural language has no separate metalanguage. See also Fodor (forthcoming).
6. See especially Harris (1982), and Harris, Gottfried, Ryckman, et al. (in press).
7. This thus cuts deeper than the naive rule-counting metrics for adjudication of grammars advocated not so long ago by generativists (see Ryckman 1986).
8. This work is reported in depth in Harris et al. (in press). These science languages occupy a place between natural language and mathematics, the chief difference from the former being that operator-argument likelihoods are much more strongly defined, amounting in most cases to simple binary selection rather than a graded scale. One of the many interesting aspects of this research is determining empirically the form of argumentation in science. The logical apparatus of deduction and other forms of inference are required only for various uses to which language may be put, rather than being the semantic basis for natural language, as has sometimes been claimed.
9. This is a refinement of the notion of distributional meaning developed in, e.g., Harris (1954).
10. The case of zero likelihood is covered by the word classes of the first constraint.
11. An example is the elision of one of a small set of operators including *appear, arrive, show up*, which have high likelihood under *expect*, in *I expect John momentarily*. The adverb *momentarily* can only modify the elided *to arrive*, etc., since neither *expect* nor *John* is asserted to be momentary. The infinitive *to*, the suffix *-ly* , and the status of *momentarily* as a modifier are the results of other reductions that are described in detail in Harris (1982).
12. For a computer implementation, see Johnson (1987). I am grateful to Tom Ryckman for helpful comments on an early draft of this review.

The Computational Analysis of English: A Corpus-based Approach

Roger Garside, Geoffrey Leech, and Geoffrey Sampson, (eds.)
(University of Lancaster and University of Leeds)

London: Longman, 1987, xii+196 pp.
ISBN 0-582-29149-6; (sb)

Reviewed by
Michael Lesk
Bell Communications Research

Why is it so remarkable to have a book whose analysis of language is entirely based on actual writing? Professors Garside, Leech, and Sampson have the refreshing view that the analysis of language ought to be based on real language, and have presented 12 papers resulting from their studies using the Lancaster-Oslo-Bergen corpus of a million words of British English. They present studies of spelling correction, part-of-speech assignment, parsing, and speech synthesis based on probability techniques derived from corpus studies. The methods here work on arbitrary texts and with reasonable efficiency.

English includes a great variety of constructions that pose a dilemma for any strict grammar: to include everything and face great ambiguity, or to be extremely prescriptive and reject much. The authors solve this problem by using probabilities to balance both frequent and infrequent constructions, and to emphasize low-level simple algorithms over deep interpretation.

For anyone trying to make practical use of text, this book is extremely enlightening. English is not an inferior substitute for Prolog, and treating it as such is not only a mismatch, but also unnecessary for many tasks. The simple use of probabilities can perform many tasks that at first glance might be thought to require understanding. Methods for doing these are explained clearly in the book.

The most detailed result described is the technique of tagging, or assigning parts of speech statistically. By using both the individual probabilities of different parts of speech for a single word, and the combined probability of sequences of two parts of speech, tagging can be done with 96–97% accuracy. This relatively simple algorithm, relying for performance on statistical data accumulated over a large sample of English rather than upon some kind of model of language, is typical of the results presented in this book. The algorithm runs on any input, from any subject area, and does a useful job without claiming to "understand" natural language. Just as we have learned that computers can play master-level chess by exhaustive evaluation of all possible moves, without any grand strategy or even plausible move selection, it seems that many linguistic tasks do not require understanding or modeling, but merely experience, translated into probability data.