DOCUMENT RESUME

ED 292 324 FL 017 241

AUTHOR	Wasow, Thomas
TITLE	Linguistics in the Study of Information and Intelligence. Linguistics in the Undergraduate Curriculum, Appendix 4-J.
INSTITUTION	Linguistic Society of America, Washington, D.C.
SPONS AGENCY	National Endowment for the Humanities (NFAH), Washington, D.C.
PUB DATE	Dec 87
GRANT	EH-20558-85
NOTE	11p.; In: Langendoen, D. Terence, Ed., Linguistics in the Undergraduate Curriculum: Final Report; see FL 017 227.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Cognitive Processes; *College Curriculum; Computer Science; Correlation; Higher Education; *Information Science; *Intelligence; *Interdisciplinary Approach; Language Processing; *Linguistics; Linguistic Theory; Logic; Philosophy; Program Descriptions; Psychology; Relevance (Education); Undergraduate Study
IDENTIFIERS	*Cognitive Sciences; Stanford University CA

ABSTRACT

Stanford University's new Symbolic Systems Program is an interdisciplinary undergraduate program focusing on understanding the nature of intelligent behavior. It brings together the disciplines of cognitive psychology, logic, computer science and artificial intelligence, philosophy, and linguistics in a newly emerging field of research concerned with the structure, content, and processing of information. Lingustics plays a central role in the program because, as the systematic study of human language, it can contribute greatly to the development of a general theory about how information is conveyed through symbols. The field of linguistics also deals with an exceptional range of phenomena, and is intellectually and practically accessible to an undergraduate student, making it an especially suitable vehicle for teaching undergraduates how to evaluate theories. (MSE)

[illegible] in the Study of [illegible] and Intelligence

by

Thomas Wasow
Stanford University

The views expressed are those of the authors and do not necessarily reflect the position of the LSA or the National Endowment for the Humanities.

The Linguistics in the Undergraduate Curriculum Project was funded by the National Endowment for the Humanities, Grant [illegible], D. Terence Langendoen, Principal Investigator.

Linguistic Society of America
1325 18th Street, N.W., Suite 211
Washington, DC 20036
(202) 835-1714

PREFACE

The Linguistics in the Undergraduate Curriculum (LUC) project is an effort by the Linguistic Society of America (LSA) to study the state of undergraduate instruction in linguistics in the United States and Canada and to suggest directions for its future development. It was supported by a grant from the National Endowment for the Humanities during the period 1 January 1985-31 December 1987. The project was carried out under the direction of D. Terence Langendoen, Principal Investigator, and Secretary-Treasurer of the LSA. Mary Niebuhr, Executive Assistant at the LSA office in Washington, DC, was responsible for the day-to-day administration of the project with the assistance of Nicole VandenHeuvel and Dana McDaniel.

Project oversight was provided by a Steering Committee that was appointed by the LSA Executive Committee in 1985. Its members were: Judith Aissen (University of California, Santa Cruz), Paul Angelis (Southern Illinois University), Victoria Fromkin (University of California, Los Angeles), Frank Heny, Robert Jeffers (Rutgers University), D. Terence Langendoen (Graduate Center of the City University of New York), Manjari Ohala (San Jose State University), Ellen Prince (University of Pennsylvania), and Arnold Zwicky (The Ohio State University and Stanford University). The Steering Committee, in turn, received help from a Consultant Panel, whose members were: Ed Battistella (University of Alabama, Birmingham), Byron Bender (University of Hawaii, Manoa), Garland Bills (University of New Mexico), Daniel Brink (Arizona State University), Ronald Butters (Duke University), Charles Cairns (Queens College of CUNY), Jean Casagrande (University of Florida), Nancy Dorian (Bryn Mawr College), Sheila Embleton (York University), Francine Frank (State University of New York, Albany), Robert Freidin (Princeton University), Jean Berko-Gleason (Boston University), Wayne Harbert (Cornell University), Alice Harris (Vanderbilt University), Jeffrey Heath, Michael Henderson (University of Kansas), Larry Hutchinson (University of Minnesota, Minneapolis), Ray Jackendoff (Brandeis University), Robert Johnson (Gallaudet College), Braj Kachru (University of Illinois, Urbana), Charles Kreidler (Georgetown University), William Ladusaw (University of California, Santa Cruz), Ilse Lehiste (The Ohio State University), David Lightfoot (University of Maryland), Donna Jo Napoli (Swarthmore College), Ronald Macaulay (Pitzer College), Geoffrey Pullum (University of California, Santa Cruz), Victor Raskin (Purdue University), Sanford Schane (University of California, San Diego), Carlota Smith (University of Texas, Austin), Roger Shuy (Georgetown University), and Jessica Wirth (University of Wisconsin, Milwaukee).

1 Introduction

The following pages sketch briefly some of the exciting new developments resulting from the collaboration of linguists with investigators from other disciplines sharing a concern with how intelligent agents process and communicate information about the world. Special attention is given to the role of linguistics in these developments, and to a new undergraduate program at Stanford University designed to train future generations of interdisciplinary researchers in this field.

Let us begin with the following rather mundane situation:

> The telephone rings, and a child answers.
>
> "Hello."
>
> "Hello. Is a grownup there?"
>
> The child calls his mother, and she picks up the telephone.

This is an extremely simple sequence of events by human standards. Now suppose that we wanted to design a machine to play the child's role. What are some of the things that the machine would have to be able to do? It would need to:

- recognize discrete words in a continuous stream of sound;
- know the meanings of individual words;
- attend to aspects of grammatical structure relevant to the meaning of what is said; for example, distinguishing the question *Is a grownup there?* from the statements *A grownup is there* and *There is a grownup*;

- take relevant contextual factors into account; for example, determining that *there* in this exchange means the location of the answerer, though in other contexts it could refer to other locations;

- on the basis of knowledge about the world, about human goals and actions, and about social conventions, infer the caller's intentions and respond appropriately—that is, get a grownup to the telephone, rather than simply giving the literally correct but clearly inappropriate response *Yes*.

The ability to conduct even the simplest conversation involves abilities that, until recently, were not associated with machines—abilities like recognizing, knowing, attending to, taking into account, and inferring. Indeed, most, if not all, aspects of what we think of as intelligence are called upon in normal, everyday language use. Hence, the study of language use is a particularly rich source of insight into the nature of intelligent behavior.

Programming a computer that genuinely understands language is, as the example illustrates, an enormously complex and difficult task. Computer scientists working on it have had some preliminary successes with specialized routines for handling very restricted types of utterances, but these tend to be difficult or impossible to extend or transport. What is needed is a theory of language use that is at once rigorous enough to be computationally implementable and flexible enough to deal with the subtleties of human language. Trying to build a language understanding system without such a theory is like trying to build a calculator on a case-by-case basis, without a theory of arithmetic.

A number of disciplines have contributed to the establishment of such a theory. Cognitive psychology provides experimental evidence for the ways in which humans perceive, classify, and reason about their environment. Logic provides mathematically sophisticated characterizations of meaning and inference for formal languages, which serve as powerful theoretical tools to apply to natural languages. Artificial intelligence provides a rapidly growing arsenal of devices for the representation and manipulation of information; while these have been developed largely for the simulation of specialized "expert" knowledge, many have useful applications as well in modeling such commonplace (but in many ways more remarkable) abilities as language understanding. Philosophy provides a tradition, over two millennia old, of careful inquiry into the nature of human knowledge and its relationship to the world. Finally, of course, linguistics plays an especially central role: it is linguistics that provides precise and detailed accounts of the sound patterns of languages (in physical, physiological, and psychological terms), as well as a rich tradition of theories and descriptive devices for the analysis of grammatical structures and their functions.

The development of a theory of language use capable of supporting a genuine language understanding technology will involve the coordinated efforts of all of these disciplines. Many promising interdisciplinary collaborations are contributing to a newly emerging field of research concerned with the structure, content, and processing of information.

2 The Role of Linguistics

For a number of reasons, linguistics plays a special role in this enterprise, and its significance will receive wider recognition as this area of investigation assumes increased technological and commercial importance in the coming decades.

Natural languages are the most highly developed symbolic systems in existence. No artificial language (including computer languages) can compare with any natural language in the variety of syntactic forms permitted, nor in the range and subtlety of meanings that can be expressed. Other naturally occurring symbolic systems (bird calls and bee dances, for example) are likewise relatively impoverished in comparison with human languages. A general theory of how information is conveyed through symbols thus can draw heavily on the systematic study of human language, that is, on linguistics.

To illustrate this point, consider the question of how the elements in a relation are differentiated in artificial languages, using the division operator as an example. Artificial languages use one of two techniques: either the arguments are given in a canonical order (e.g., $12 \div 3 = 4$), or they are identified with keywords (e.g., *dividend: 12, divisor: 3, quotient: 4*). Each of these strategies has its advantages: the former is notationally compact, whereas the latter allows the elements to be introduced in any order. There are analogues to both of these formal devices in natural languages: English uses word order to differentiate subject from object (*The man saw the woman* vs. *The woman saw the man*), whereas Japanese uses particles adjacent to the nouns:

otoko	ga	on'na	o	mita
man	SUBJ	woman	OBJ	saw

"The man saw the woman"

on'na	ga	otoko	o	mita
woman	SUBJ	man	OBJ	saw

"The woman saw the man"

It is the particles *ga* and *o* that indicate who did the seeing and who was seen; reversing the order of the nouns would not alter this. Thus, *on'na o otoko ga mita* also means "The man saw the woman."

In addition to these two strategies, however, natural languages have others that serve the same general purpose. In Russian, for instance, the roles of the participants in a sentence are indicated by changes in the form of the nouns in the sentence. Thus, we have

chelovek	videt	zhenshchinu
man	sees	woman

"the man sees the woman"

zhenshchina	videt	cheloveka
woman	sees	man

"the woman sees the man"

Again, the word order is not essential, nor are there distinct particles to mark the difference between subject and object. Finally, some languages mark the verb, rather than the nouns, to indicate who did what to whom. In Abkhaz, a language of the Caucasus, one would say:

a-	xàc'a	a-	pħ°ə̀s	də-y-bè-yt'
the	man	the	woman	her-he-sees

"The man saw the woman"

Here it is the form of the verb that indicates that it is the man who sees the woman, rather than vice versa. As in Japanese and Russian, the order of the nouns is not important.

As the above examples illustrate, natural languages exhibit a wide range of formal devices for conveying information, including some that have not been exploited in artificial languages. Natural languages provide a rich source of ideas about the ways in which information can be encoded in symbols. Each of the different strategies illustrated above serves the same general purpose, but they may well differ with respect to such matters as learnability and processing difficulty. Designers of artificial languages might well learn something from a closer look at natural languages.

Moreover, anyone interested in natural language processing by computers quickly comes to realize that failure to attend to apparently arbitrary grammatical details will, in the long run, lead to misunderstandings. The phenomenon of subject-verb agreement in English, for example, appears at first glance to be completely redundant, since the singular/plural distinction is marked both on the noun and the verb. This has led some builders of natural language processing systems to believe that it could be ignored: the number marked on the subject would be used to determine the semantics, agreement would not be checked, some ill-formed input (e.g., *The men is here*) would be accepted, and no harm would be done. Eventually, however, this strategy is doomed to failure, for even such seemingly meaningless bits of grammar as agreement serve to resolve ambiguities in some cases. If, for example, a sales executive were to tell the company's customer database *List every company with Japanese affiliates that buys widgets*, the answer would very likely not be the same as the answer to *List every company with Japanese affiliates that buy widgets.* A natural language interface that failed to distinguish these two sentences could cost a company millions of dollars. Only linguists have detailed theories of such apparently arcane facts about language structure, so designers of natural language systems need training in linguistics.

Linguistics is exceptional, too, in the range of phenomena it deals with. Under the umbrella of linguistics fall such diverse aspects of language as the physical properties of speech sounds, the physiology of the organs of speech and hearing (including the relevant parts of the brain), the patterns of regularities exhibited by related word forms, the grammatical patterns of languages, the meanings of words, how word meanings are combined into phrase meanings and sentence meanings, the relationship between literal and conveyed meaning, the variations of pronunciation and syntax across groups of speakers and circumstances of use, and how languages change. Thus, linguistics is concerned with all facets of one symbolic system (natural language), from its medium to its message, from its forms to its functions.

No other discipline looks at any symbolic system from such a variety of perspectives. This is important in part because of the subtle ways in which the information conveyed can be affected by the form in which it is expressed. For example, the stress pattern in a sentence like *John insulted Bill after he criticized Mary* can affect the reference of the pronoun: if the verbs *insulted* and *criticized* are stressed, *he* will be interpreted as John, but if *he* gets heavy stress, *he* will be taken as referring to Bill. Only by attending to diverse aspects of the system can such interactions be analyzed. Certainly, any hope of developing fully automatic speech understanding systems will depend on having theories broad enough to deal with facts like this.

In short, anyone concerned with how information is conveyed and processed should know something about how natural languages are structured and used. Linguistics offers a wealth of theoretical concepts for the analysis of sentence structure and linguistic sound patterns, developed over a period of twenty-five centuries. In the decades since World War II, there has been an explosive growth in this discipline, resulting in theories of unprecedented precision and generality. The electronics revolution that has occurred in the same period has also created powerful new tools for the analysis and synthesis of speech. It is just beginning to have a similar impact on other areas of linguistics.

One final attribute of linguistics that is of interest in the present connection is its accessibility. Despite its long history and theoretical sophistication, most of modern linguistics is comprehensible to an intelligent undergraduate. Unlike the physical sciences, in which current research questions can only be understood after years of study, the frontiers of linguistics are accessible after only a few courses. There are several reasons for this, two of which deserve special mention. First, most areas of linguistics depend less heavily on complex mathematical results than is common in many other fields; hence, extensive mathematical training is not a prerequisite to doing advanced work in linguistics. (Work on speech synthesis and analysis, cited above, is an exception). Second, since every normal human is a native speaker of a natural language, we all have a rich store of (typically unsystematized) knowledge about language prior to any formal study, a store that can be tapped to permit students to make very rapid progress in understanding how natural language works.

One very concrete way in which everyone's tacit knowledge of language serves linguistics instruction is as a source of data. While other sciences require the student to become familiar with elaborate laboratory techniques that will, with considerable effort on the student's part, produce data relevant to the formulation and testing of hypotheses, linguists can perform crucial experiments merely by concocting strings of words and assessing their well-formedness. This can often be done instantaneously, without leaving one's seat. Hence, experiments in linguistics can be performed in class, without any special equipment, and, in many cases, by the student. This makes it possible for the teacher to concentrate on argumentation and theory development, rather than on techniques of data collection. The result is very rapid progress to the frontiers of the field. Consequently, it is common in linguistics for undergraduates to do original research, in some cases even publishable research.

Linguistics, then, is a particularly suitable vehicle for teaching undergraduates how to evaluate theories by drawing out their empirical consequences and designing test cases. It gives them the opportunity to experience first-hand what it is like to formulate hypotheses, evaluate them experimentally, and write up the results. This facilitates the development of valuable thinking and writing skills that should be applicable to a wide variety of other endeavors. Hence, linguistics would be a useful component in almost any student's undergraduate education. For the reasons given earlier, it is a must for any student primarily interested in questions concerning information and intelligence.

3 Stanford's Symbolic Systems Program

Because of the many points of contact between linguistics and other aspects of the study of information and intelligence, it is evident that the development of a general theory of language will, in the long run, depend on the next generation of researchers, whose multidiscplinary training must begin early in their careers. Towards this end, Stanford University has recently initiated a new undergraduate major, called Symbolic Systems.

The program requires study in five traditional disciplines: Computer Science, Linguistics, Logic, Philosophy, and Psychology. Each student must complete a common set of eleven core courses in these fields, plus a concentration in one of eight areas: artificial intelligence, cognitive science, computation, logic, natural language, philosophical foundations, semantics, or speech. It is excellent preparation for graduate study in any of several fields, or for employment in the information industry.

Stanford is an ideal setting for the establishment of such a program, for it has long played a leading role in the study of information and intelligence. With world class departments of computer science, linguistics, philosophy, and psychology, it has a long history of interactions among these fields. An interdisciplinary research program in cognitive science was established in the late 1970s, with funding from the Alfred P. Sloan Foundation. More recently, a gift from the System Development Foundation led to the establishment of the Center for the Study of Language and Information (CSLI), a unique institution that brings together scholars from academia and researchers from industry, all concerned with problems of language and information.

Its founders hope that Stanford's leadership in these areas of research will give the Symbolic Systems Program a high degree of visibility, which

will lead, in turn, to the establishment of similar programs at other cclleges and universities.