A metric for assessing the complexity of semantic (and pragmatic) analysis in natural language processing is proposed as part of a general applied theory of linguistic semantics for NLP. The theory is intended as a complete projection of linguistic semantics onto NLP and is designed as an exhaustive list of possible choices among strategies of semantic analysis at each level, from the word to the entire text. The alternatives are summarized in a chart, which can be used for existing or projected NLP. The remaining components of the applied theory are also outlined.

1. Goal
The immediate goal of the paper is to explore the alternative choices in the analysis of meaning in natural language processing (NLP). Throughout the paper, semantics subsumes pragmatics. The more ambitious goal of the paper, however, is to lay ground for an applied theory of linguistic semantics for NLP (ALT/NLP).

2. Applied Theory of Linguistic Semantics for Natural Language Processing
ALT/NLP is a part of an applied linguistic theory for natural language processing (ALT/NLP). The latter obviously includes other components, most prominently syntax and morphology. The applied theory is the result of a projection of linguistic theory onto the NLP plane or, in other terms, an adaptation of general linguistic theory specifically for NLP purposes.

2.1. Linguistic Theory, Semantic Theory. The modern concept of linguistic theory, developed primarily by Chomsky (1965), is that of a set of statements which (i) characterizes language as a complex structure and describes that structure top down, 2) underlies each description of a particular language and determines the format of such a description, Semantic theory as part of linguistic theory determines semantic descriptions. Semantic descriptions assign meanings to sentences, and each meaning is a formula logically deduced from the rules provided by semantic theory and utilized in the description. A valid semantic description assigns each sentence the same meaning that the native speaker does.

The intended use of ALT/NLP is to bring to the NLP customer, not necessarily knowledgeable in linguistics, the totality of what linguistics knows about meaning by (i) listing all the choices available at each level of semantic analysis, 2) determining causal connections among choices and the propagation of constraints through the choice space, 3) assessing any existing NLP system as to the complexity of its semantic equipment and the possibilities of expanding it in the desired direction if necessary, and 4) relating each chain of compatible choices to the practical needs and resources. This paper deals almost exclusively with the first item on this agenda.

3. The Complexity Scale of Semantic Analysis
The scale proposed in this section is a list of choices available at each of the five levels of semantic analysis corresponding to the five meaningful linguistic entities pertinent to NLP - the word, the clause, the sentence, the paragraph, and the text, or discourse. At each level, attention is paid to such dimensions as the completeness and relative depth of analysis.

All the examples are taken from one paragraph (i) in Ullman (1982:1-2). The paragraph does not stand out in any sense except that it clearly belongs to the computer sublanguage of English.

(i) (i) Data, such as the above, that is stored more or less permanently in a computer we term a database.
(ii) The software that allows one or many persons to use and/or modify this data is a database management system (DBMS).
(iii) A major role of the DBMS is to allow the user to deal with the data in abstract terms, rather than as the computer stores the data.
(iv) In this sense, the DBMS acts as an interpreter for a high-level programming language, ideally allowing the user to specify what must be done, with little or no attention on the user's part to the detailed algorithms or data representation used by the system.
(v) However, in the case of a DBMS, there may be a far less relationship between the data as seen by the user and as stored in the computer, than between, say, arrays as defined in a typical programming language and the representation of those arrays in memory.
3.1. The Word. The semantic descriptions of the words are usually stored in the dictionary of an NLP system. The analysis at the word level may be full or partial. The analysis in full if every word of the analyzed text is supposed to have a non-empty (i.e., distinct from just the spelling) entry in the dictionary. The analysis is partial if only some words must have an entry. Thus, an analysis of (11) as a sequence of three key words (for instance, in automatic abstracting), as shown in (2), is definitely partial.

(2) DATA COMPUTER DATABASE

The analysis may be limited or unlimited. The analysis is unlimited if the meaning of the word needs to be utilized in its entirety. The analysis is limited if, for the purposes of a given NLP, it would suffice, for instance, to describe the words in (3i) as physical objects and the words in (3ii) as mental objects and cut off all the other elements of their meanings.

(3) (i) person, operator, computer

Another version of limited analysis would be to analyze the meanings of the words to the point of distinguishing each word from any other word and no further. Thus, operator and computer can be distinguished in terms of semantic description as shown in (4).

(4) (i) operator: Physical Object, Animat
(ii) computer: Physical Object, Inanimate

It is worth noting that while person and operator can be similarly distinguished along lines of (4), they cannot be distinguished in the computer sublanguage and are, therefore, complete synonyms. In other words, person is the parent of operator in English as a whole but not in this sublanguage.

(5) (i) person: human
(ii) operator: human, Using Gadget

The analysis can use a number of methods. The first and simplest case seems to be the net-membership, e.g., key-word analysis. Within this approach, words are assigned to certain semantic classes, represented by what is often called key words or descriptors, and this remains their only characteristic. In more sophisticated versions, descriptors may be further subcategorized, e.g., parent-child relations among them can be set up, and dictionary entries will then contain hierarchies of them, e.g., (6).

(6) data MINERAL COURT COMPUTER-RELATED

Second, a form of feature (or componential) analysis can be used. The main distinction between feature analysis and net-membership is that, in the former, features come from different hierarchies. Thus, for (6) to be an example of feature analysis rather than of descriptor analysis, COMPUTER RELATED should not be a child of MENTAL OBJECT in the system.

Third, the dictionary entries may be set up as networks. In linguistic semantics, the concept of semantic field (e.g., for instance, Resnik 1980:31-2) corresponds to a primitive network. In a pure network-based approach, actual words serve as the nodes - there are no set-members or categorical markers (unlike in syntactical trees) and no primes (unlike in feature analysis). The networks may have weighted or unweighted links (edges); they may also, or alternatively, be labeled or unlabeled. The number of labels may vary. The labels can also be set up as the other kind of nodes. Generally, the nodes can be equal (flat) or unequal (hierarchical). Thus, red may be set up as a node while red in a slot of a physical object, connected with the redness node by the link color.

3.2. The Clause. The clause boundaries are obtained through the application of a syntactic parser. The full/partial dimension of this level deals with whether every clause of the sentence is analyzed or none are emitted, and the latter is not impossible. The unlimited/limited dimension deals with the detailedization of the analysis along the various parameters (see below). Decisions on both of the dimensions may be predetermined by those taken at the word level. In general, the full/partial and unlimited/limited dimensions become the more trivial or obvious along the higher level. Accordingly, while fully reflected at each level on the chart in (10), they will be hardly mentioned in the subsequent sections.

The most important decision to make at the clause level is whether the output is structured or not. The unstructured output will simply list the semantic characteristics of all the words in the clause which have them, in the order of their appearance. The only clause-related information in such a case will be the clause boundaries.

The structured output may be dependent on the natural-language syntax of the clause or not. The accepted term is: semantic interpretation for syntactically-dependent outputs, and semantic representation, otherwise. In a typical semantic representation, a tree-like structure, such as (10) (cf. Nirenburg et al. 1985:229), may be set up for clauses instead of their regular syntactic structures, with the nodes and/or link labels being of a different nature. An event with its actors as in (7i) should be an obvious possible choice for the analysis of the clause. The structures may be more or less distant from the syntactic structure (in any guise) but the presence of just one semantic node or - more often - link label would render them non-syntactic.

(7) (i) [data is stored more or less permanently in the computer]
(iii) store agent object time space goal

In (7ii), the deviations from syntactic structure abound and include most prominently 1) different link labels, e.g., goal; 2) substitution of sublanguage-determined paraphrases, e.g., always for more or less permanently; 3) information not contained in the clause and supplied from the sublanguage knowledge base, e.g., goal = maintain-database.

Whether information for the semantic analysis of the clause is supplied from outside for its analysis or only from inside determines whether the analysis is supercompositional or compositional. Finally, the clause analysis may include or exclude superpropositional information. Exclusively propositional analysis will basically analyze the clause as a sentence. Thus, (7i) will be analyzed without the square brackets around data, which signify that the word is the supplied antecedent for a proximal entity (that). Superpropositional analysis typically subsumes propositional analysis and adds to it the information on the links of the clause with the other clauses of its own and/or the adjacent sentences. Thus, in the case of (7ii), that should be related to data two clauses earlier and the nature of the link should be described: syntactically, i.e., as a relative clause; however, a semantic label, such as PROVISION, would be much more informative (see also below).

3.3. The Sentence. The first important phenomenon to consider at the sentence level is whether the sentence is represented as a clausal discourse structure or not. If the sentence is not represented as such a structure, it becomes simply a sequence of clauses augmented by syntactical dependency information. Such a sequence will not be much distinct from a sequence of monoclusal sentences, except that some of them will be clustered together. If the clausal discourse structure is there, it will be probably presented as a graph with the clauses for nodes and relations between them for link labels. Again, as in the case of the clause, the link labels may range from the syntactic terms to semantic relations. A more semantically informative structure, with semantic link labels, is illustrated in (8) for (11):

339
Data... we term a database (0)

Repsion such as the above that is stored more or less permanently in a computer

Semantic Link labels are often associated with non-syntactic clauses being distinguished - thus, such as the above is not a full-fledged syntactic clause.

Like clause analysis, sentence analysis may be compositional or supercompositional. There is such supercompositional information available at this level than at the clause level. The supercompositional information is, of course, knowledge-based. It can include 1) semantic field information for words (paradigmatic semantic information), i.e., that computer in (1) is a machine or a mechanical device and that certain other words, probably not in the sublanguage, are fellow members of the field; 2) information on the relations of the sentence with the world or subworld (for a sublanguage), e.g., for (1), the meaning of each sentence is clarified if semantic analysis utilizes a rule about the subworld, namely that every mental object in the subworld is located in the computer memory; 3) speech act information, i.e., whether the sentence is an assertion, a question, a command or any other possible value of the illocutionary-force variable (see Nirenburg et al. 1985:1234); 4) information on the links of the sentence with other sentences (see the next paragraph); 5) given/new information, e.g., that this data is given in (iii); (vi) main clause information.

Information on the links of the sentence with other sentences includes connectives, both explicit as, for instance, however in (iv), and implicit. This information is crucial for establishing the discourse structure of the paragraph (see 3.4).

Information is used only in systems which accommodate extraneous information and ignore by systems with exclusively semcompositional information.

Finally, each sentence can be characterized as to the goal it expresses. In a textbook exposition like (i), the goal tends to be amotivational - it is to convey information or to teach, but in a narrative text with protagonists or in a dialogue, goals can vary with each cue (see Schank and Abelson 1977; Reichman 1985).

3.4. The Paragraph. The semantic analysis of the paragraph may include its representation as a sentential discourse structure or not include it. If there is no such representation, then similarly to sentence analysis, the paragraph will be treated simply as a linear sequence of sentences. Otherwise, the paragraphs may be represented as a graph with sentences for nodes and with relations between the sentences for label links. No standard syntactical nomenclature is available for this level. Using one simple semantic link label, (i) may be represented as

(9)

\( (\text{Expansion}) \)
\( (\text{iii}) \)
\( (\text{iv}) \)

Because of the nature of (i) and of its sublanguage, the links between the sentences are such less diverse than in a casual discourse - and this is good for NLP. It is possible, and often advisable to combine the clause structures of the sentences and the sentential structures of the paragraph in one graph, because frequently a clause in one sentence is linked to a clause in another rather than the whole sentence to the other, and the resulting graph is more informative.

It is also important to decide at this level whether to develop paragraph topic extraction or not. For the former option, the paragraph can be summarized by creating a new sentence or, alternatively, one of the existing sentences is selected to "represent" the whole paragraph.

3.5. The Text. The questions of paraagraph discourse structure and of textual topic extraction arise here similarly to paragraph analysis.

4. A Semantic Metric for NLP.

(10) summarizes all the main options for semantic analysis in NLP (P-level).

(10) Semantic Metric for NLP:

| WORD | CLAUSE | SENTENCE | PARAGRAPH | TEXT |
|------|--------|----------|------------|------|
| Full | Full   | Full     | Full       | Full |
| Limited | Limited | Limited | Limited     | Limited |
| Method | :Comp. | :CLBound. | :SenBound. | :ParBound. |
| set/sea- | Prop. | :Disc.Str. | :Disc.Str. | :Disc.Str. |
| ture/act | :Comp. | :Topic Extr. | :Topic Extr. | :Goal |
| vi. |

Each system of NLP can use (10) to chart out its own method of semantic analysis, both before and after its formulation, and to compare itself with any other system (the actual metric is derived from (10) by adding an obvious measure of distance). Naturally, there are four possible basic types of semantic analysis in NLP than add, simply because many values in (10) determine others and render many combinations incompatible. On the other hand, there are variations within the basic types.

The proposed metric is just one part of ALTL/SLP. The complete ALTL/SLP adds the following parts to the metric: 1) mutual determination and exclusion of values in (10); 2) choices for execution of each value; 3) relations between NLP needs and values and combinations of values.

It should be noted that besides ensuring the total modularity of semantic analysis in NLP by providing the full/partial and unlimited/limited values for each level, this part of the theory is itself modular in the sense that any value or option, which may have been left out inadvertently or which may emerge in the future, can be added to (10) without any problem.

5. References:

Chomsky, N. 1965. Aspects of the Theory of Syntax. Cambridge, MA: M.I.T. Press.

Kittredge, R. and J. Lehrberger 1982. Sublanguage: Studies of Language in Restricted Semantic Domains. Berlin - New York: de Gruyter.

Nirenburg, S. (ed.) 1985. Proceedings of the Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages. Hamilton, N.Y.: Colgate University.

Nirenburg, S., V. Raskin, and A. R. Tucker 1985. "Interlingua design for TRANSLATOR." In: Nirenburg (1985), pp. 230-40.

Raskin, V. (W.) 1971. К теории языковых подсистем /Towards a Theory of Linguistic Subsystems/. Moscow: Moscow University Press.

Raskin, V. 1974. "A restricted sublanguage approach to high quality translation." American Journal of Computational Linguistics 11:3, 5-10.

Raskin, V. 1983. A Concise History of Linguistic Semantics. W. Lafayette, IN: Purdue University, 3rd ed.

Raskin, V. 1980c. "Linguistic and encyclopedic information in text processing," Quaderni di Semantica VI:1, pp. 92-102.

Raskin, V. 1983b. "Linguistics and natural language processing." In: Nirenburg (1985), pp. 286-92.

Reichman, R. 1985. Seeing the Computer to Talk Like You And Me. Cambridge, MA: M.I.T. Press.

Schank, R. and R. Abelson 1977. Scripts, Plans, Goals, and Understanding. Hillsdale, N.J.: Lawrence Erlbaum.

Schank, R., L. Birenbaum, and J. Mey 1985. "Integrating semantics and pragmatics," Quaderni di Semantica VII:2, pp. 313-26.

Ullman, J. D. 1982. Principles of Database Systems. Rockville, MD: Computer Science Press, 2nd ed.