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 completed for each existing or projected NLP system. 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 enhances 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). 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 1) 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 theoretical inadequacy of much of contemporary linguistics may stem from Chomsky's view that the theory is one. An alternative view of theory as the exhaustive list of alternatives, complete with the issues on which the alternatives differ and the consequences of each choice, is simply indispensable for applications.

2.2. Linguistic Applications and NLP. A meaningful application of linguistics always deals with a problem which comes entirely from the area of application and not from linguistics. Every NLP system requires the description of a natural language fragment, often of a sublanguage. In the one hand, modern linguistics, with its emphasis on formality, would seem to be uniquely and unprecedentedly qualified to supply such a description. On the other hand, while every single fact about language the NLP expert needs is out there in linguistics, much of it is not easily accessible. Descriptions posing as theories or theories posing as descriptions tend not to list all the necessary facts in any way facilitating computer implementation (see below). The only solution to the problem is to develop a general and systematic way of projecting linguistic knowledge onto NLP, which is what ALT/NLP is all about.

2.3. Applied Theory, II: ALT/NLP. ALT/NLP deals with pretty much the same facts and phenomena of language as linguistics per se. There are, however, crucial differences. First, while both "pure" and "applied" theories are formal, the nature of the formalism is different. Second, pure linguistic theory deals with a language as a whole while ALT/NLP deals with limited and relatively closed sublanguages or language fragments (see Resnik 1971, 1974, 1985; Kittredge and Lehrberger 1982).

Third, pure linguistic theory must ensure a complete and even coverage of everything in the texture of language; ALT/NLP analyzes only as much as needed for the purposes of NLP and ignore all the linguistic information that is superfluous for it. Fourth, the ultimate criterion of validity for pure linguistic theory is the elusive explanatory adequacy; the ultimate criterion for ALT/NLP is whether NLP systems resulting from its application work.

Fifth, pure linguistic theory can afford not to pursue the issue once a method or a principle is established. In ALT/NLP, everything should be done explicitly to the very end, and no extrapolation is possible. And finally, pure linguistic theory has to be concerned about the boundary between linguistic and encyclopedic knowledge, i.e., between our knowledge of language and our knowledge of the world (cf. Resnik 1985a). There may be no particular need to maintain this distinction in an NLP system (cf. Schank et al. 1965) because the computer needs all the kinds of available information for processing the data.

2.4. Applied Theory, II: ALT/NLP. ALT/NLP, a projection of linguistic semantics onto NLP, is designed to serve all the various NLP systems. Therefore, it is viewed and set up as the exhaustive list of possibilities for semantic analysis and description available in linguistic semantics.

The intended uses of ALT/NLP is to bring to the NLP customer, not necessarily knowledgeable in linguistics, the totality of what linguistics knows about meaning by 1) 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 (1) in Biber (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 far less relationship between the data as seen by the user and as stored in the computer, than between, say, scenes 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 is 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 (1(i)) 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 (3(i)) as physical objects and the words in (3(ii)) as mental objects and omit all the other elements of their meanings.

(3) (i) person, operator, computer
(ii) data, database, algorithm

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, Animatable
(ii) computer: Physical Object, Inanimate

It is worth noting that while person and operator can be similarly distinguished along the lines of (5), (ii), 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: Machine, Using Gadget

The analysis can use a number of methods. The first and minimal one seems to be the net-membership approach, 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, i.e., parent-child relations among them can be set up, and dictionary entries will then contain hierarchies of them, e.g., (6).

(6) data  MINERAL  CURVE  COMPUTER-RELATED

Second, a form of feature (or compositional) analysis can be used. The main distinction between feature analysis and net-membership is that, in the former, the 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 MINERAL CURVE in the system.

Third, the dictionary entries may be set up as networks. In linguistic semantics, the concept of semantic field (see, for instance, Resnik 1984:31–2) corresponds to a primitive network. In a pure network-based approach, only actual words serve as the nodes — there are no metasets 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 node can be equal (flat) or unequal (hierarchical). Thus, redness may be set up as a node while red is 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 distinction at this level deals with whether every clause of the sentence is analyzed or some are omitted, and the latter is not impossible. The unlimited/limited dimension deals with the detailization 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 and obvious along the clause level. Accordingly, while fully reflected at each level on the chart in (10), they will be only mentioned in the subsequent subsections.

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 terms are: 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. 1985:125), 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 arguments as in (7(i)) 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]
(ii) store agent object time space goal

In (7(i)), the deviations from syntactic structure abound and include most prominently 1) different link labels, e.g., goal; 2) substitution of sub-language-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 semantical analysis of the clause is supplied from outside the clause as well as from inside its analysis or only from inside determines whether the analysis is suprapositional or compositional.

Finally, the clause analysis may include or exclude suprapositional information. Exclusively suprapositional analysis will basically analyze the clause as a sentence. Thus, (7) will be analyzed without the square brackets around data, which signify that the word is the supplied antecedent for a pronoun entity (that).

Suprapositional analysis typically subsumes suprapositional 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 (7(i)), that should be related to data two clauses earlier and the nature of the link should be described: syntactically, it is a relative clause; however, a semantic label, such as EXPANSION, 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 clauseal 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 monoclausal sentences, except that some of them will be clustered together. If the clauseal 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 (1):
4. A Semantic Metric for NLP.

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

(10) Semantic Metric for NLP:

<table>
<thead>
<tr>
<th>WORD</th>
<th>CLAUSE</th>
<th>SENTENCE</th>
<th>PARAGRAPH</th>
<th>TEXT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full</td>
<td>Full</td>
<td>Full</td>
<td>Full</td>
<td>Full</td>
</tr>
<tr>
<td>Limited</td>
<td>Limited</td>
<td>Limited</td>
<td>Limited</td>
<td>Limited</td>
</tr>
</tbody>
</table>


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 many possible basic types of semantic analysis in NLP that are suitable, 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 ALT/NLP. The complete ALT/NLP 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/unlimited 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:


