filename: a:\sentactics\ws2\taxonomy

SENTACTICS
A Natural Language Parsing System for English

DEFINITIONS

The definitions and classification of language units which follow are based on a largely non-traditional and non-transformational approach to parsing. It is therefore necessary to re-define significant linguistic units and to reclassify them according to some radical changes in form and function. In some cases, familiar terms are used in traditional ways, in others, familiar terms are re-defined in non-traditional ways, and, finally, some new terms are employed.

The SENTAX approach to parsing bears more than a passing resemblance to the immediate constituent analysis of traditional structural linguistics. It does not depend, however, on the binary "cuts" from the whole sentence down to single components but moves, rather, from the constituent relationships of single adjacent units through intermediate levels to an ultimate realization of the entire sentence as a unit. It deals, further, with the written (or printed) representation of English resulting in the consideration of written spacing and punctuation conventions rather than dealing with the intonation patterns and other suprasegmental features of English speech.

Thus the following definitions and classification:

Word: a basic unit of meaning (or meanings) bounded by spaces.

Words are classified into the following basic categories:

Form class:	adverb [adv]
	adjective [adj]
	noun [N]
	verb [V]
Structure:	determiner [D]
	preposition [P]
	relative pronoun [R]
	subordinate conjunction [S]
	intensifier [inten]
	interjection [intj]

modal [mod]
verb marker [V_m]
auxiliary [aux]
verbal [Vb] (past participle [V-en], present participle [V-ing], infinitive [to+V])

Word cluster: a group of two or more adjacent words belonging to the same word class, occurring in a prescribed order, with left or right-branching internal constituency.

noun cluster [N_{cl}]

adjective cluster [adj_{cl}]

determiner cluster [D_{cl}]

preposition cluster [P_{cl}]

Compounds: a group of two or more adjacent words belonging to the same word or unit class, with "0" internal constituency.

Compounds are ultimately treated by connecting with "and" or comma.

Word group: a group of two or more adjacent words belonging to different word classes. Word groups may incorporate word clusters, considered as the equivalent of a single word.

Noun group [N_{gr}] a group of adjoining words consisting of a prenominal unit, a pronominal unit and a postnominal unit.

Prenominal: D, D_{cl}, or D_0 [null]
adj, adj_{gr}, V-ing, V-en, adj_0, or $preN_0$

Pronominal: N_{hw} or noun surrogate (adj[$_{gr}$] V-ing[$_{gr}$] V-en[$_{gr}$], P_{gr})

Postnominal: adv, P_{gr}, R_{gr}, V-ing, V-ing_{gr}, V-en, V-en_{gr}, to+V, to+V_{gr}, or $posN_0$

sentactics

Verb group [V_{gr}]: a group of words containing a preverbal unit, a proverbal unit and a postverbal unit.

preverbal: adv, V_m, mod, aux, aux_{gr} or $preV_0$

proverbal: V_{hw}

postverbal: adj, adj_{gr}, N_{gr}, P_{gr}, R_{gr}, S_{gr} or $posV_0$

Adjective group [adj_{gr}]: int + adj

Preposition group [P_{gr}]: $P_{[cl]}$ + N_{gr}

Relative group [R_{gr}]: R + Vb_{gr}, R + sentence

Subordinate group [S_{gr}]: S + P_{gr}, S + $Vb_{[gr]}$ S + sentence

Verbal group [Vb_{gr}]: V-en, V-ing or to+V + adv, $adj_{[gr]}$, or N_{gr}.

Predication: a group of two or more words in which one word (or word group) says something about another word (or word group).

Primary predication: N_{gr} + V_{gr} with any necessary word form required by agreement/tense conventions.

Secondary predication: words or word groups in a postnominal position.

Tertiary predication: words, word clusters or word groups in a prenominal position.

Sentence: a group of words, word clusters, or word groups, bounded by an initial and terminal "0" and containing at least one primary predication, stated or presupposed.

Entry point: the initial word in any word group. In structure word groups, the structure word itself is the entry point. In any N_{gr} or V_{gr}, the entry point is a word other than the noun or

sentactics

verb (or "$\underline{n}_0$", to designate a structural position for the entry point).

Constituency: the perceivable structural/semantic relationship between any adjacent units (words, word clusters or word groups).

Right-branching constituency: the first of two adjacent units "pointing to" the second unit, indicated by a right-pointing arrow [-->].

Left-branching constituency: the second of two adjacent units "pointing to" the first, indicated by a left-pointing arrow [<--].

Mutual constituency: Each of the two adjacent units pointing to the other, indicated by a double-headed arrow [<-->].

Null constituency: no perceivable relationship between two adjacent units, indicated by "null" [0]. (This degree of constituency is relative rather than absolute--a case could be made quite often that there is a relationship, albeit remote, between adjacent words or word groups. But the ultimate test comes when a much stronger relationship can be clearly marked between larger units with each of the two adjoining words belonging to a separate unit.)

GENERAL RULES

Constituency is determined through a series of levels until there is a series of words bounded by an initial and terminal "0". These "0's" are sentence boundaries. To determine all the constituent relationships within those boundaries, the following rules must be observed:

sentactics

All sentences formed for the purpose of conveying information are of the basic form N_{gr} + V_{gr}. In such sentences the normal entry point for the sentence is the first word of the N_{gr}. (Other possible sentence entry points and their implications for structural change will be shown and discussed later.)

At the first level, there will be a null ("0") placed before the first word, either because it is the initial word in the unit to be analized or because there is little perceivable relationship between it and the preceding word.

Then, constituency must be determined between each set of adjacent words. In a sentence, for instance, consisting of words "a, b, c, d, e, f," sets would be comprised of "ab," "bc", "cd", etc. To determine constituency between adjacent words, relationships must be determined as right-branching, left-branching, mutual or "0".

In the N_{gr}, constituency is resolved from "0" to branching constituency within the prenominal by starting at the word or cluster immeddiately to the left of the N_{hw} and working back at successive levels to the entry point.

It is necessary, quite often, to go beyond the particular set of words and to determine the relationship of the second word to its following adjacent word. It is necessary in the case of a v-ing following a D, for instance. The v-ing is either a modifier in the prenominal or a noun surrogate as the pronominal. If the word following the v-ing is a N, the v-ing relationship to the N is, "v-ing --> N" which makes the initial relationship, "D 0 v-ing". In the n_{gr}, "The dancing bear," the constituent relationship would thus be, "the [0] dancing [-->] bear."

sentactics

But in the construction, "the dancing always amazes me," the "0" relationship between "dancing" and "always"* indicates that "dancing" is a noun surrogate. The constituent relationship between "the" and "dancing" would thus be, "the [-->] dancing [0] always." This analysis confirms the traditional distinction between the present participle in a participial (adjectival) function or in a gerundive (nominal) one.

First-level constituency in the Noun-group:

Entry point and prenominal:

The entry point for any N_{gr} is $d[_{cl}]$. If no $D[_{cl}]$ word appears as the entry point, the entry point must be identified as D_0. This creates a limitation on the classes of nouns or noun surrogates which may then appear in the N_{hw} position.

If the entry point (D, D_{cl} or D_0) is followed by int or adj, the constituent relationship must be described as, "$D[_{cl,0}]$ 0 [int,adj]". In a few instances, governed by the word or words following, the adj will be revealed as a noun surrogate, in which case the relationship between D and adj would be shown as, "D --> adj".

If it is determined that there is a string of adj's, it becomes necessary to distinguish between a compounding of adj and adj_{cl}. In a few instances (such as " . . . little old lady . . . ") the adj's are from subclasses that demand a given order. When this is not the case, the adj's are considered to be compounded: that is, they are considered to be connected by "$and_{[0]}$" or $comma_{[0]}$.

---

*This relationship will be covered in more detail later.

sentactics

With no intervening words the constituent relationship will be designated as "$D[_{c1,0}]$ --> Nhw."

The next word will be one of the following:

D (indicating that the entry point is actually a D_{c1}).

int, adj, v-ing, v-ed, N

N_{hw}

If D, indicate constituent relationship as mutual (<-->). If next word is D, indicate the relationship as <-->; continue until the following word is not D.

If int, adj, v-ing, v-ed, indicate constituent relationship as null (0)