Tagging as a Means of Refining and Extending Syntactic Classes

Catherine Macleod, Adam Meyers, and Ralph Grishman
Computer Science Department
New York University
{macleod,meyers,grishman}@cs.nyu.edu

1 Comlex Syntax

Comlex Syntax is a moderately-broad-coverage English lexicon (with about 38,000 root forms) being developed at New York University under contract to the Linguistic Data Consortium; the first version of the lexicon was delivered in May 1994. The lexicon is available to members of the Linguistic Data Consortium for both research and commercial applications. It was developed for use in processing natural language by computer.

Comlex Syntax is particularly detailed in its treatment of subcategorization (complement structures). It includes 92 different subcategorization features for verbs, 14 for adjectives, and 9 for nouns. These distinguish not only the different constituent structures which may appear in a complement, but also the different control features associated with a constituent structure.

In order to make this dictionary useful to the entire NLP community, an effort has been made to provide detailed yet theory neutral syntactic information. In part, this involved using categories that are generally recognized, i.e. nouns, verbs, adjectives, prepositions, adverbs, and their corresponding phrasal expansions np, vp, adjp, pp, advp. COMLEX cites the specific prepositions and adverbs in prepositional and particle phrases.[1]

We selected as a starting point, the classes for complements and features developed by the New York University Linguistic String Project (LSP) [2], since the coverage is very broad and the classes well defined. We augmented and further refined these classes by studying the coding employed by several other major lexicons used for automated language analysis. We consulted the the Oxford Advanced Learner's Dictionary (OALD) [3], the Longman Dictionary of Contemporary English (LDOCE) [4] , the verb codes developed for English by Sanfillipo as part of the ACQUILEX project[7], and The Brandeis Verb Lexicon.[2] We adopted a Brandeis-like notation for COMLEX complement names. The Brandeis notation is compositional, consisting of lists of elements joined by hyphens e.g. p1-ing-sc (preposition "about" followed by a gerund where the matrix subject is the subject of the gerund e.g. "he lied about going"). In adapting this notation for COMLEX, we fixed the list of complement names and then provided a separate explicit definition of the syntactic structure associated with each complement name. Further information on these classes and definitions can be found in our Reference Manual [8] and COMLEX Word Classes Manual [9].

2 Tagging Task

We are in the midst of tagging a corpus with COMLEX syntactic classes. We are tagging verb complements (as defined in COMLEX) for 100 examples each of 750 common verbs which have previously been entered in the COMLEX lexicon. The original motivation for this was twofold, (1) to gather statistics on the frequency of occurrence of a particular complement of a verb and (2) to check on our coverage, ascertaining that we had not missed the most commonly occurring complements.

The corpus consists of Brown (all, i.e. 7 MB), Wall Street Journal (27 MB), San Jose Mercury (30 MB), Associated Press (29.5 MB), Miscellaneous (Treebank literature 1.5 MB) etc. adding up to about 100 MB of text.

In creating the initial COMLEX, we have followed traditional dictionaries in classifying verbs by the complements with which they can appear *in isolation* in simple declarative sentences. This classification is certainly useful in understanding the argument structure of the verbs.

However, this approach runs into conflict with the task of tagging examples in a corpus. Complements may be transformed (sometimes beyond ready recognition) or contextually zeroed. The

[1] e.g. pp :pval "to" for *he went to the party*; part-np :adval "up" for *he woke up the child.*

[2] Developed by J. Grimshaw and R. Jackendoff.

complement may occur in the same sentence as in topicalization and passivization, it may be zeroed but recoverable (to a greater or lesser degree) as in wh-clauses, and wh-questions, it may be zeroed and recoverable semantically or it may be zeroed but recoverable only from discourse analysis or it can be ambiguous. To be consistent with our original approach, we reconstruct the complements where possible. Furthermore, we have noted that not all verbs are equally subject to particular types of contextual zeroing of complements.[3] Although we are not certain yet how to classify the verbs based on this information, we are being careful to record the type of zeroing involved in tagged examples so that this information can later be recovered and studied.

3 Passivization and Topicalization

The recovery of the complement in passivization and topicalization is reasonably straight-forward, though passivization may lead to misinterpretation of the complement. In a sentence like (1) given the distance between "order" and "to dig", the tendency is to mark the to-infinitive as part of the complement rather than part of the noun phrase. A similar case is (2). In examples (3) - (4) the separated pp's and np are, in fact, part of the COMLEX complement. [4]

(1) Orders were GIVEN to dig. [np]
(2) Annual authorizations of $15 million were ADDED for area vocational education programs that meet national defense needs for highly skilled ..[np]
(3) sets were developed and distributed, and lantern slide teaching sets on 21 pathology subjects were ADDED to the loan library of the Medical Illustration Service. [np-pp :pval 'to']
(4) The front part of my head was CALLED a face, and I could talk with it. [np-np-pred]
(5) To that Rousseau could AGREE. [pp :pval 'to']
(6) Even if that's all the promise he ever GAVE... [np]
(7) Arthur Williams had to be located, they AGREED. [that-s]

Topicalization examples (5) through (7) show that the complement is readily accessible. However, even here we can see that in example (7) the complement appears to need a that-complementizer when it occurs after the verb. Topicalization does not allow a that-complementizer

(8)*That Arthur Williams had to be located, they agreed.

so we either have to state that "agree" takes a bare sentence or we have to add material that is not in the text.[5].

4 Wh-clauses

The existence of "missing" complements has put us into the uncomfortable position of tagging items that do not appear in the text. If the complement can be recovered straightforwardly from the surrounding sentence, we mark the verb for that complement. For example, in relative clauses the complement can usually be recovered.

(9) .... to sit more patiently with what they have BOUGHT. [np]
(10) There is perhaps no value statement on which people would more universally AGREE than the statement that intense pain is bad. [pp :pval 'on']
(11) ''What have you GOT on today''? she inquired. [part-np :adval 'on']
(12) Where were they all WALKING to? [pp :pval 'to']
(13)I know where to GO to buy a cheap coat. [where :no-class t].

In all the above cases, except for sentence (13)[6] the complement can be unambiguously recovered. In sentence (9) they bought something, in (10) they would agree on the statement, and in (11) he/she has got something on. However, even though "where" is to be reconstructed in both (12) and (13) only in (12) can it be unambiguously interpreted as a pp (they were walking to somewhere), in (13) "where" could be interpreted as a pp or an advp (go there/go to some place) so we classify it as having a non-comlex-class "where".

5 Parentheticals

Further "missing" complements were found in parentheticals.

[3] e.g "He suggested that I should go and I agreed." zeroed 'to go' vs *"He suggested that I should go and I wanted." where it is not possible to zero 'to go'.

[4] NB in the examples the capitalized verb is the one in question. Unless otherwise specified these examples are all from the concordance.

[5] However, see further on, the question raised about obligatory *that-s*

[6] Example (13) is not from the concordance

(14) For example, to move (as the score REQUIRES) from the lowest ˜F-major register up to a barely audible ˜N minor in four seconds, not skipping, at the same time, even one of the 407 fingerings, seems a feat too absurd to consider, and it is to the flautist's credit that he remained silent throughout the passage.

(15) The ideal home, they AGREED, would be a small private house or a city apartment of four to five rooms, just enough for a family

Reconstructing a *to-inf* for "requires" in (14) would not be correct since "require" needs a *np-to-inf* (*the score requires to move), but it is not clear what the np could be (perhaps "the flautist"? the tone?). We felt these cases to be different from the other cases that we have discussed above not only because of the difficulty of locating the complement but in the nature of the construction. This construction is more similar, in fact, to our *v-say* feature which allows a verb like "say" to occur in sentence adjunct positions without its complement.[7]

(16) He said, ''I want to see you.''

(17) ''I,'' he said, ''want to see you.''

(18) ''I want'', he said, ''to see you.''

(19) ''I want to see you,'' he said.

Therefore, we concluded that the fact that these verbs can occur without their complements is a fact about the grammar of parentheticals and not about the individual verb and is therefore not a proper consideration of the dictionary. These examples, then, we are tagging as "parenthetical" which is not a COMLEX class . However, the information that these verbs can occur in parenthetical constructions will be contained in the tags.

6 The "Intransitive" Question

We have encountered several types of zeroing in the corpus and have established new COMLEX classes of intransitives to take them into consideration. These classes were established to deal with the observed intransitivity of verbs which ordinarily are not seen as intransitives. For example, in isolation, "agree" may not occur intransitively unless it has a plural subject (our intrans-recip class). (these examples are not from the concordance)

(20) he agreed with her.

(21) they agreed. (with each other)

(22) *he agreed.

However, the data is rife with examples of intransitive "agree" occurring with a singular subject as seen by the following examples.

(23) the gourmet insisted that it is done that way at the most fashionable dinners, the girl reluctantly AGREED.

(24) Why, it's all right, isn't it, Mother''? Her woolly-minded parent AGREED.

''Of course, dear'', she said. ''It's only that I like to know where you go''.

(25)''He's one hell of a decent boy. I like that kid''. ''I AGREE, yes''.

(26)... he hoped to persuade him to become his assistant in research for the labor novel; if Breasted AGREED, they would get a car and tour the country,

(27)..spoke up, ''plenty of it. Let me give Papa blood''. The doctor AGREED, but explained that it would be necessary first to check Fred's blood to ascertain whether or not it was of the same type...

We have established the COMLEX class *intrans-ellipsis* for these cases and since we feel that the complement is "underlyingly" present (the tagger is able to supply the missing material) we would like to be able to reconstruct a complement for the above instances of "agree". There seem to be two possibilities: (A) where someone agrees with someone that-s (in (23) she agreed [with him/that it was done that way],[8] in (24) she agreed [with her/that it was all right], in (25) I agree [with you/with that/that he is a decent boy]); (B) a to-infinitive (in (26) if he agreed [to become his assistant], in (27) he agreed [to let he/she give him blood]).

Even though this last example (27) presents some difficulties in reconstruction (1) because it occurs outside the sentence containing the verb and (2) because there is a change of mood from imperative to infinitival, we can understand that the doctor agreed to let [him] give blood and reconstruct a subject controlled to-infinitive. The COMLEX entry will be

```
(INTRANS-ELLIPSIS :SUBC (to-inf-sc))
```

The others (sentences 23-25) would be tagged, arbitrarily, as having a prepositional phrase contain-

[7] Examples not from concordance.

[8] There is also the reading "she agreed to do it that way"

ing the preposition "with" and they will be entered in the dictionary with the COMLEX class *INTRANS-ELLIPSIS :SUBC (pp :pval ("with"))*. *Intrans-Ellipsis*, therefore, will differentiate between "true" intransitives [9] and cases like the above.

We also found occurrences of "habitual" intransitives in the text. Even verbs which are always considered to be transitive, like "hit" for example, can be used intransitively if the action is considered to be habitual. [10]

(28) That child always hits.
(29) She always abbreviates, a very annoying habit.
(30) He nagged constantly.

We are have made this another class of intransitives and are tagging them as

[INTRANS :NOT-IN-DICT ''habitual'']

in order to gather statistics.[11] Since it seems that this is really a grammatical question, as any verb (it would seem) may occur as a habitual intransitive, we would consider adding *intrans-habitual* to COMLEX as a complement class only for verbs which are often used in this way.

7 Tagging Improves COMLEX

Aside from presenting us with these interesting here-to-fore unthought of phenomena, tagging has also tightened up our classification of some complements, leading us in the direction of combining some that had been separate and re-grouping others. Among verbs that take sentential complements, we have made a distinction between those verbs that require the that-complementizer and those that do not. In our tagging, we have found that many verbs, which COMLEX has classified as requiring *that-s*, appear in the texts with a bare sentence. It is too early to tell whether to remove this distinction but from what we have seen so far this seems a possible outcome. Similarly, we have a frame-group which classes together a number of wh-complements (we made the distinction between wh- and how). Now it looks as if we will need to make a different grouping with "whether"/"if" contrasting with "what" and "how".

All in all, our tagging has been unexpectedly interesting and informative. We are acquiring not only statistical data on the occurrence of complements in texts but information on possible gaps in COMLEX's syntactic coverage which we can move to rectify, if it seems justified or, at least, we will have a record in our tagged data. We have often been asked why we are not machine tagging instead of painstakingly hand tagging. I think our response now is obvious, with machine tagging we would not have been able to recognize and record these facts about language.

8 Future Directions

We have performed a few experiments with our initial tags, both to evaluate their quality and to determine whether we can learn something about the syntax-semantics interface.

Another resource readily available to the Natural Language Processing community is WordNet. WordNet is a repository of information about the lexical semantics of English which has been under development at Princeton University, under the direction of George Miller, and his group [5], [6] for nearly a decade. One of the aims of WordNet is to provide broad coverage of English vocabulary; the current version includes over 80,000 words and collocations.

WordNet is also tagging the Brown Corpus, in their case with sense tags. Unlike COMLEX, WordNet intends to tag only the Brown Corpus but with senses for all the major parts of speech.

One of the interesting possibilities for a multiply tagged corpus of this type is to try to ascertain whether the syntactic and the semantic classes line up in a significant way and whether it is possible to use the syntactic classes to disambiguate among senses of a word. Another possibility is to see if one can succeed in collapsing word senses by considering regular correspondences between these senses and the observed complement structures.

In [1], we showed that the comparison of COMLEX's syntactic classes and the word senses of WordNet produced interesting results for two verbs *know* and *remain*. COMLEX classes were found to be perfectly aligned with some senses and good predictors of others, leading one to posit that perhaps the latter senses might be candidates for melding into one sense. We have begun a few further experiments based on the premise that to some degree semantic classes of verbs are predictable from syntactic subcategorization.

[9] e.g. sleep, in *he slept* and arrive in *he arrived*

[10] Examples not from concordance

[11] We also use intrans-habitual to refer to generic situations as well, e.g. "As a group, three year old children, hit."

References

[1] Catherine Macleod, Ralph Grishman, and Adam Meyers Developing Multiply Tagged Corpora for Lexical Research *To appear in Proc. of the Post COLING International Workshop on Directions of Lexical Research*, Beijing, China, August, 1994.

[2] Eileen Fitzpatrick and Naomi Sager. The Lexical Subclasses of the LSP English Grammar Appendix 3. In Naomi Sager *Natural Language Information Processing.* Addison-Wesley, Reading, MA, 1981.

[3] A. S. Hornby, editor. *Oxford Advanced Learner's Dictionary of Current English.* 1980.

[4] P. Proctor, editor. *Longman Dictionary of Contemporary English.* Longman, 1978.

[5] George Miller (ed.), WordNet: An on-line lexical database. In *International Journal of Lexicography* (special issue), 3(4):235-312, 1990.

[6] George Miller, Claudia Leacock, Randee Tengi, and Ross Bunker. A semantic concordance. In *Proceedings of the Human Language Technology Workshop*, pages 303–308, Princeton, NJ. March 1993. Morgan Kaufmann.

[7] Antonio Sanfilippo. LKB encoding of lexical knowledge. In T. Briscoe, A. Copestake, and V. de Pavia, editors, *Default Inheritance in Unification-Based Approaches to the Lexicon.* Cambridge University Press, 1992.

[8] Catherine Macleod and Ralph Grishman. COMLEX Syntax Reference Manual.

[9] Susanne Rohen Wolff, Catherine Macleod and Adam Meyers. COMLEX Word Classes Manual