The Influence of Tagging on the Classification of Lexical Complements

Catherine Macleod, Adam Meyers, and Ralph Grishman,
Computer Science Department
New York University
715 Broadway, 7th Floor
New York, NY 10003
{macleod,meyers,grishman}@cs.nyu.edu

Abstract

A large corpus (about 100 MB of text) was selected and examples of 750 frequently occurring verbs were tagged with their complement class as defined by a large computational syntactic dictionary, COMLEX Syntax. This tagging task led to the refinement of already existing classes and to the addition of classes that had previously not been defined. This has resulted in the enrichment and improvement of the original COMLEX Syntax dictionary. Tagging also provides statistical data which will allow users to select more common complements of a particular verb and ignore rare usage. We discuss below some of the problems encountered in tagging and their resolution.

1 Introduction

COMLEX Syntax is a moderately-broad-coverage English lexicon (with about 38,000 root forms) developed at New York University under contract to the Linguistic Data Consortium; the first version of the lexicon was delivered in May 1994. The tagged version was delivered in August 1995. The lexicon is available to members of the Linguistic Data Consortium for both research and commercial applications. It was developed specifically 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 Natural Language Processing community, an effort was 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 associated with particular verbs.[1]

As a starting point, the classes for complements and features developed by the New York University Linguistic String Project (LSP) (Fitzpatrick, 1981), were selected since the coverage is very broad and the classes well defined. These classes were augmented and further refined by studying the coding employed by several other major lexicons used for automated language analysis. The Oxford Advanced Learner's Dictionary (OALD) (Hornby, 1980), the Longman Dictionary of Contemporary English (LDOCE) (Proctor, 1978), the verb codes developed for English by Sanfilippo as part of the ACQUILEX project(Sanfilippo, 1992), and The Brandeis Verb Lexicon [2] were consulted. A Brandeis-like notation was adopted 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, the list of complement names became fixed and a separate explicit definition of the syntactic structure associated with each complement name was provided. Further information on these classes and definitions can be found in the Reference Manual (Macleod, 1993) and the COMLEX Word Classes Manual (Wolff, 1995).

2 Tagging Task

We tagged 100 examples for each of 750 common verbs which had previously been entered in the COMLEX lexicon. These tags, which became part of the dictionary entry, contain the location of the example and the name of the verbal complement identified at that location (see figure 1 for a sample

[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.

tagged entry). The original motivation for tagging was twofold, (1) to gather statistics on the frequency of occurrence of a particular complement of a verb and (2) to check on COMLEX coverage, ascertaining that the most commonly occurring complements had not been overlooked in the original entries.

The corpus used for this tagging consists of Brown (all, i.e. 7 MB), Wall Street Journal (17 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.

The tags in figure 1 are all from the Brown Corpus. We chose to give preference to tagging examples from Brown. In that way, we could overlap, as much as possible, with other tagging efforts that have been done on the Brown Corpus by the Penn Treebank and WordNet.

In the creation of the original COMLEX, traditional dictionary procedure was followed by classifying verbs as having 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 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 the original approach, the complements have been reconstructed where possible. Furthermore, we have noted that not all verbs are equally subject to particular types of contextual zeroing of complements.[3] The type of zeroing involved in the examples has been recorded in the tags and added to the dictionary.

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. In examples (2) - (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-pp :pval "for"]
(3) sets were developed and distributed,
    and lantern slide teaching sets on 21
    pathology subjects were ADDED to the
    loan library of the Medical Illustra-
    tion 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 locat-
    ed, 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.

4 Wh-clauses

The existence of "missing" complements forces one 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, the verb was marked 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 univer-
      sally 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"]
```

[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. The complement with which the verb should be tagged appears in square brackets. Unless otherwise specified these examples are all from the corpus.

```
(VERB :ORTH "adjust"
      :SUBC (((P-POSSING :PVAL ("to"))
              (PP :PVAL ("for" "to"))
              (NP-PP :PVAL ("for" "to"))
              (NP)(INTRANS))
      :TAGS ((TAG :BYTE-NUMBER 6602672
                  :SOURCE "brown"
                  :LABEL (NP))
              (TAG :BYTE-NUMBER 6203043
                  :SOURCE "brown"
                  :LABEL (PP :PVAL ("to")))
              (TAG :BYTE-NUMBER 5717471
                  :SOURCE "brown"
                  :LABEL (NP))
              (TAG :BYTE-NUMBER 5537823
                  :SOURCE "brown"
                  :LABEL (NP-PP :PVAL ("to")))
```

Figure 1: Partial Comlex Syntax dictionary entry for ADJUST.

```
(13) .. School where their delicate
     transformation BEGAN. [where]
```

In all the above cases, except for sentence (13) 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 being part of a pp (they were walking to somewhere); in (13) "where" could be interpreted as a pp or an advp (it began there/it began at school) so we classify it as having the class "where" (which is not a COMLEX complement).

5 Parentheticals

Further "missing" complements were found in parentheticals.

```
(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 remain-
     ed 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 COMLEX's V-SAY feature which allows a verb like "say" to occur in sentence adjunct positions without its complement.[5]

```
(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. These examples, then, have been tagged as "parenthetical" and the new COMLEX feature PARENTHETICAL has been given to the verbs which can occur in parenthetical constructions.

6 The "Intransitive" Question

We have encountered several types of zeroing in the corpus which occur with verbs which we would normally consider transitive[6] or verbs which can be intransitive only under special circumstances. For example, in isolation, "agree" may not occur intransitively unless it has a plural subject (COMLEX intrans-recip class).[7]

```
(20)  he agreed with her.
(21)  they agreed. (with each other)
```

[5] Examples not from the corpus.

[6] Note, that while for some "transitive" verbs are defined only as verbs which take NP complements, we consider verbs "transitive" that take any type of complement, including PPs

[7] These examples are not from the corpus

(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 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 him/her 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 tag entry is

(INTRANS-ELLIPSIS :ELLIP (to-inf-sc))

Intrans-ellipsis is the name of the class and what is elided (:ellip) is a subject controlled to infinitive (to-inf-sc, a COMLEX complement). The others (sentences 23-25) were tagged, arbitrarily, as having a prepositional phrase containing the preposition "with" and they will be entered in the dictionary with the tag INTRANS-ELLIPSIS :ELLIP (pp :pval ("with")). The new COMLEX complement INTRANS-ELLIPSIS is added to verbs of this type and therefore COMLEX differentiates 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 tagged these [INTRANS-HABITUAL].[11] Since it seems that this is really a grammatical question, as any verb (it would seem) may occur as a habitual intransitive, it has not been proposed as a COMLEX complement.

7 New Noun Phrase Complement, NADVP

During our tagging, we found that there are a group of noun phrases that pattern with adverbs and prepositional phrases, which we have called NADVP's (Noun Adverbial Phrases). These are divided into NADVP-TIME, NADVP-DIR, NP-NADVP-LOC and NADVP-MANNER. Often these expressions are adjuncts, but in the following examples, they are complements since they are required to produce a grammatical sentence.(The examples in this section are not from the corpus but similar examples were found)

The meeting took 3 hours. [nadvp-time]
*The meeting took.
He headed home/east/that way. [nadvp-dir]
*He headed.
He put the stakes every five feet.
[np-nadvp-loc]
*He put the stakes.
He put it that way. [np-nadvp-manner]
*He put it.

These noun phrases may be substituted for by adverbs or prepositional phrases.

The meeting took 3 hours. [nadvp-time]
The meeting took long. [advp]
He headed home/east/that way. [nadvp-dir]
He headed to the store. [pp]
He put the stakes every five feet.
[np-nadvp-loc]
He put the stakes at designated places.
[pp]

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

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

[10] Examples not from the corpus

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

```
He put it that way. [np-nadvp-manner]
He put it firmly. [nadvp-dir]
```

In general these verbs do not take regular np complements, at least not with the same meaning.

```
The meeting took/lasted 3 hours.
He took/*lasted the car.
He headed/went home.
He headed/*went the cow down the road.
*He put the stakes the table.
*He put it the interest.
```

8 NUNITP: to Tag or not to Tag

Another class of noun phrases caused us great soul searching. A number of verbs take very particular noun phrases. Verbs like "increase", "decrease" and "expand" take complement groups which require a noun with the subclass NUNIT. [12] These verbs occurred predominately in environments like

```
The price increased 5 to 10 percent.
The price increased 5 dollars a share.
```

We decided not to make this a separate NP complement for several reasons: (1) these verbs also take regular NP complements, though in some instances (as in the below example) the meaning of the verb changes. As COMLEX does not sense disambiguate the semantic difference does not affect the dictionary entry.

```
"Those vitamins increased his appetite."
```

(2)the NUNITPs are not syntacticly distinguished; other nouns occur with similar structures.

```
"He ate 5 to 10 pickles (a day)."
```

On the other hand, the increase-type verbs can appear with a whole range of nunitp complements (complements which contain an nunitp[13]):

```
The price increased (5%) to $10 (a share).
The price increased (5%) from $10 (a share).
The price increased from $10 (a share) to
$30 (a share).
The price increased to 30 dollars from 10
dollars.
The price increased by 5% to end at $100.
```

whereas verbs like "eat" can not

```
*He ate to/from 10 pickles.
*He ate by 10% to 20 pickles (a day).
```

Although we decided not to add NUNITP as a separate NP complement, we have let the NUNITP tags for verb complements remain, to reflect the information that in our corpus this type of verb occurs almost exclusively with this type of NP. We have added a separate frame group with the name NUNITP-TO-RANGE which includes the complements mentioned above. Although, it is called NUNITP to underline the fact that ordinarily the nouns that occur are NUNITs or are coerced into being NUNITs in this structure[14], the NPs are not formally distinguished as such in the notation of the frame group. The fact that these noun phrases, and the NADVPs above, behave in a manner distinct from other NPs is recognized and discussed in Ross's paper on Defective Noun Phrases (Ross, 1995)

9 Tagging Improves COMLEX

Aside from presenting these interesting and unexpected phenomena, tagging has tightened up the classification of some complements, leading in the direction of combining some complements that had been separate and re-grouping others. COMLEX had a frame-group which classified together a number of wh-complements. Now there is a different grouping with the original "whether" / "if" / "what" (WH-S complement) and "how" (HOW-S, PP-HOW-TO-INF) augmented by "where" / "when" / "how much" / "how many" (WHERE-WHEN-S). This last group was established for verbs like "define" and "forecast" which do not take members of the original frame groups.

```
 "Last year, the Supreme Court DEFINED
  when companies, such as  military
  contractors, may defend themselves."
*The Supreme Court defined if companies
  may defend themselves.
 "Ptolemy's problem is to FORECAST
  where, against the inverted bowl of
  night, some  particular light will be
  found at future times."
*The problem is to forecast how to find
  the light.
?The problem is to forecast how he will
   react.
```

Tags that were not deemed worthy to become COMLEX complements for various reasons (e.g. rarity or sublanguage use) are defined in the COMLEX Syntax Manual for Tagged Entries (Meyers, 1995). All in all, our tagging has been interesting and informative. We have acquired not only statistical data on the occurrence of complements in texts but information on possible gaps in COMLEX's syntactic coverage which we moved to rectify, when it seemed justified, and we have a record in our tagged data of those instances which we did not add to COMLEX classes. We have

[12] These are nouns which can appear in quantifier phrases including a scalar adjective before another noun or as a head noun followed by a prepositional phrase containing a scalar noun (a two FOOT long board/a board two FEET in length).

[13] The nunitp is $/dollars in the examples

[14] Compare "The price increased by five percent to a total of 2,000 dollars per share." "The contents of each barrel increased by 5 pickles to a total of 25 pickles per barrel."

often been asked why we did not machine tag instead of painstakingly hand tagging. We think our response now is obvious, with machine tagging we would not have been able to recognize and record these facts about language.

Acknowledgements

Design and preparation of COMLEX Syntax has been supported by the Advanced Research Projects Agency through the Office of Naval Research under Awards No. MDA972-92-J-1016 and N00014-90-J-1851, and The Trustees of the University of Pennsylvania.

References

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.

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

Catherine Macleod and Ralph Grishman. COMLEX Syntax Reference Manual. New York University, 1993

Adam Meyers, Catherine Macleod and Ralph Grishman. COMLEX Syntax Manual for Tagged Entries. New York University, 1995

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

John Robert Ross. Defective Noun Phrases. In Barbara Need et.al., editors, *Papers from the 31st Regional Meeting of the Chicago Linguistic Society.* To appear.

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.

Susanne Rohen Wolff, Catherine Macleod and Adam Meyers. COMLEX Word Classes Manual New York University, 1995