The Comlex Syntax Project: The First Year

Catherine Macleod, Ralph Grishman, and Adam Meyers

Computer Science Department
New York University
715 Broadway, 7th Floor
New York, NY 10003

ABSTRACT

We describe the design of Comlex Syntax, a computational lexicon providing detailed syntactic information for approximately 38,000 English headwords. We consider the types of errors which arise in creating such a lexicon, and how such errors can be measured and controlled.

1. Goals

The goal of the Comlex Syntax project is to create a moderately-broad-coverage lexicon recording the syntactic features of English words for purposes of computational language analysis. This dictionary is being developed at New York University and is to be distributed by the Linguistic Data Consortium, and to be freely usable for both research and commercial purposes by members of the Consortium.

In order to meet the needs of a wide range of analyzers, we have included a rich set of syntactic features and have aimed to characterize these features in a relatively theory-neutral way. In particular, the feature set is more detailed than those of the major commercial dictionaries, such as the Oxford Advanced Learner’s Dictionary (OALD) [4] and the Longman Dictionary of Contemporary English (LDOCE) [8], which have been widely used as a source of lexical information in language analyzers. In addition, we have aimed to be more comprehensive in capturing features, and in particular subcategorization features, than commercial dictionaries.

2. Structure

The structure of COMLEX has been discussed at length in our report to the 1993 HLT Workshop so we will briefly touch on the details of our dictionary entry. The major classes (adjectives, nouns and verbs) are marked for features and complements (subcategorization frames), examples of which can be seen in Figure 1.

Nouns have 9 possible features and 9 possible complements; adjectives have 7 features and 14 complements; and verbs have 5 features and 92 complements. Figure 2 shows some actual dictionary entries, including some entries for adverbs and prepositions.

In order to insure the completeness of our codes, we studied the coding employed by several other major lexicons, including the Brandeis Verb Lexicon², the ACQUILEX Project [10], the NYU Linguistic String Project [9], the OALD, and the LDOCE, and, whenever feasible, have sought to incorporate distinctions made in any of these dictionaries. The names for the different complement types are based on the conventions used in the the Brandeis Verb Lexicon. The notation indicates the type and order of the elements (NP = noun phrase, PP = prepositional phrase, NP-PP = a noun phrase followed by a prepositional phrase, pval = the selected prepositions).

The subcategorization types are defined by frames. These frames which appear in our reference manual (see Figure 3) include the constituent structure :es, the grammatical structure :gs, optional :features and one or more examples :ex. The features in the subcategorization frames are not those in the dictionary but refer to the control or raising properties of the verb where applicable. In particular, they capture four different types of control: subject control, object control, variable control, and arbitrary control. Furthermore, the notation allows us to indicate that a verb may have different control features for different complement structures, or even for different prepositions within the complement. We record, for example, that “blame ... on” involves arbitrary control (“He blamed the problem on going too fast.”), whereas “blame for” involves object control (“He blamed John for going too fast.”).

There are two complements represented by the frames in Figure 3, possing and ing-sc. possing stands for a frame group which includes two frames *possing (where the subject of the gerund is present) and *ing-ac (where the subject is interpreted to be arbitrary).

A verb which is assigned possing must be able to occur in both of these frames. ing-sc also stands for a frame group. It includes be-ing-sc and *possing. Here the subject of the gerund must be the same as the surface subject and the possessive subject of *possing will be coreferential with the surface subject.

3. Methods

Our basic approach is to create an initial lexicon manually and then to use a variety of resources, both commercial and corpus-derived, to refine this lexicon. Although methods have been developed over the last few years for automatically identifying some subcategorization constraints through corpus analysis [2,5], these methods are still limited in the range of distinctions they can identify and their ability to deal with low-frequency words. Consequently we have chosen to use manual entry for creation of our initial dictionary.

The entry of lexical information is being performed by four graduate linguistics students, referred to as elves (“elf” = enterer of lexical features). The elves are provided with a menu-based interface coded in Common Lisp using the Garnet GUI package, and running on Sun workstations. This interface also provides access to a large text corpus; as a word is being entered, instances of the word can be viewed in one of the windows. Elves rely on citations from the corpus, definitions and citations from any of several printed
Noun feature NUNIT: a noun which can occur in a quantifier-noun measure expression
ex: "two FOOT long pipe"/"a pipe which is two FEET in length"
Noun complement NOUN-THAT-S: the noun complement is a full sentence
ex: "the assumption that he will go to school (is wrong.)"
Adj feature ATTRIBUTIVE: an adjective that occurs only attributively (ie before the
noun) and never predicatively (after "be")
ex: "The LONE man rode through the desert"/"the man was lone."
Adj complement ADJ-FOR-TO-INF: includes three infinitival complements of adjectives
ex: "it is PRACTICAL for Evan to go to school." (extrap-adj-for-to-inf)
"the race was easy for her to win." (extrap-adj-for-to-inf-np-omit)
"Joan was kind to invite me." (extrap-adj-for-to-inf-rs)
Verb feature VMOTION: a verb which occurs with a locative adverbial complement.
ex: "he ran in" (which may permute to "in he ran.")
Verb complement: NP a verb which takes a direct object noun phrase.
ex: "he ran a gambling den."

Figure 1: Some features and complements.

Figure 2: Sample COMLEX Syntax dictionary entries.

Figure 3: Sample COMLEX Syntax subcategorization frames.
dictionaries and their own linguistic intuitions in assigning features to words.

Entry of the initial dictionary began in April 1993. To date, entries have been created for all the nouns and adjectives, and 60% of the verbs\(^3\); the initial dictionary is scheduled for completion in the spring of 1994.

We expect to check this dictionary against several sources. We intend to compare the manual subcategorizations for verbs against those in the OALD, and would be pleased to make comparisons against other broad-coverage dictionaries if those can be made available for this purpose. We also intend to make comparisons against several corpus-derived lists: at the very least, with verb/preposition and verb/particle pairs with high mutual information [3] and, if possible, with the results of recently-developed procedures for extracting subcategorization frames from corpora [2,5]. While this corpus-derived information may not be detailed or accurate enough for fully-automated lexicon creation, it should be most valuable as a basis for comparisons.

### 4. Types and Sources of Error

As part of the process of refining the dictionary and assuring its quality, we have spent considerable resources on reviewing dictionary entries and on occasion have had sections coded by two or even four of the elves. This process has allowed us to make some analysis of the sources and types of error in the lexicon, and how they might be reduced. We can divide the sources of error and inconsistency into four classes:

1. **errors of classification:** where an instance of a word is improperly analyzed, and in particular where the words following a verb are not properly identified with regard to complements or different types. Specific types of problems include misclassifying adjuncts as arguments (or vice versa) and identifying the wrong control features. Our primary defenses against such errors have been a steady refinement of the feature descriptions in our manual and regular group review sessions with all the elves. In particular, we have developed detailed criteria for making adjunct/argument distinctions [6].

A preliminary study, conducted on examples (drawn at random from a corpus not used for our concordance) of verbs beginning with "j", indicated that elves were consistent 93% to 94% of the time in labeling argument/adjunct distinctions following our criteria and, when they were consistent in argument/adjunct labeling, rarely disagreed on the subcategorization. In more than half of the cases where there was disagreement, the elves separately flagged these as difficult, ambiguous, or figurative uses of the verbs (and therefore would probably not use them as the basis for assigning lexical features). The agreement rate for examples which were not flagged was 96% to 98%.

2. **omitted features:** where an elf omits a feature because it is not suggested by an example in the concordance, a citation in the dictionary, or the elf's introspection. In order to get an estimate of the magnitude of this problem we decided to establish a measure of coverage or "recall" for the subcategorization features assigned by our elves. To do this, we tagged

\(^3\)No features are being assigned to adverbs or prepositions in the initial lexicon.

The "Complements only" is the percentage of instances in the corpus covered by the subcategorization tags assigned by the elves and does not include the identification of any prepositions or adverbs. The "Complements only" would correspond roughly to the type of information provided by OALD and LDOCE.\(^4\) The "Complements + Prepositions/Particles" column includes all the features, that is it considers the correct identification of the complement plus the specific prepositions and adverbs required by certain complements. The two columns of figures under "Complements + Prepositions/Particles" show the results with and without the enumeration of directional prepositions.

We have recently changed our approach to the classification of verbs (like "run", "send", "jog", "walk", "jump") which take a long list of directional prepositions, by providing our entering program with a P-DIR option on the preposition list. This option will automatically assign a list of directional prepositions to the verb and thus will save time and eliminate errors of missing prepositions. Figure 5 shows the dictionary entry for "jump", taken from the union of the four elves. If you note the large number of directional prepositions listed under PP (prepositional phrase), you can see how easy it would be for a single elf to miss one or more. The addition of P-DIR has eliminated that problem.

In some cases this approach will provide a preposition list that is a little rich for a given verb but we have decided to err on the side of a slight overgeneration rather than risk missing any prepositions which actually occur. As you can see, the removal of the P-DIRs from consideration improves the individual elf scores.

The elf union score is the union of the lexical entries for all four elves. These are certainly numbers to be proud of, but realistically, having the verbs done four separate times is not practical. However, in our original proposal we stated that because of the complexity of the verb entries we would like to have them done twice. As can be seen in Figure 6, with two passes we succeed in raising individual percentages in all cases.

We would like to make clear that even in the two cases where our individual lexicographers miss 18% and 13% of the complements, there was only one instance in which this might have resulted in the inability to parse a sentence. This was a missing intransitive. Otherwise, the missed complements would have been analyzed as adjuncts since they were a combination of prepositional phrases and adverbials with one case of a subordinate conjunction "as".

We endeavored to make a comparison with LDOCE on the measurement. This was a bit difficult since LDOCE lacks some complements we have and combines others, not always consistently. For instance, our PP roughly corresponds to either L9 (our PP/ADVP) or prep/adv + T1 (e.g. "on" + T1) (our PP/PART-NP) but in some cases a preposition is mentioned but the verb is classified as intransitive. The straightforward comparison has LDOCE finding 73% of the tagged

\(^4\)LDOCE does provide some prepositions and particles.
| elf # | Complements only | Complements + Prepositions/Particles without P-DIR | using P-DIR |
|-------|------------------|-----------------------------------------------|-------------|
| 1     | 96%              | 89%                                           | 90%         |
| 2     | 82%              | 63%                                           | 79%         |
| 3     | 95%              | 83%                                           | 92%         |
| 4     | 87%              | 69%                                           | 81%         |
| elf av| 90%              | 76%                                           | 84%         |
| elf union| 100%              | 93%                                           | 94%         |

Figure 4: Number of subcategorization features assigned to "j" verbs by different elves.

\[\text{verb} \quad :\text{orth} \quad \text{"jump"} \quad :\text{subc} \quad ((\text{pp} \quad :pval \quad ("up" \quad "around" \quad "along" \quad "across" \quad "at"
  "down" \quad "in" \quad "from" \quad "into" \quad "through"
  "out" \quad "off of" \quad "past" \quad "over" \quad "out of"
  "onto" \quad "off" \quad "on" \quad "under" \quad "towards"
  "toward" \quad "to"))
  (\text{pp-pp} \quad :pval \quad ("about" \quad "from" \quad "on" \quad "off of" \quad "off"
  "onto" \quad "to"))
  (\text{np-pp} \quad :pval \quad ("through" \quad "over" \quad "to")) \quad (\text{intrans}) \quad (\text{np})
  (\text{part-pp} \quad :\text{adv} \quad ("up" \quad "down" \quad "off" \quad "back" \quad "away" \quad "out")
  :pval \quad ("on" \quad "from" \quad "to"))
  (\text{part} \quad :\text{adv} \quad ("off" \quad "on" \quad "across" \quad "aside" \quad "down" \quad "back"
  "away" \quad "in" \quad "up")))
  :\text{features} \quad ((\text{vmotion})))\]

Figure 5: Dictionary entry for "jump" showing proliferation of pvals.

| elf # | Complements only | Complements + Prepositions/Particles without P-DIR | using P-DIR |
|-------|------------------|-----------------------------------------------|-------------|
| 1 + 2 | 100%             | 91%                                           | 93%         |
| 1 + 3 | 97%              | 91%                                           | 92%         |
| 1 + 4 | 96%              | 91%                                           | 91%         |
| 2 + 3 | 99%              | 89%                                           | 90%         |
| 2 + 4 | 95%              | 79%                                           | 86%         |
| 3 + 4 | 97%              | 85%                                           | 92%         |
| 2-elf av| 97%              | 88%                                           | 91%         |

Figure 6: Number of subcategorization features assigned to "j" verbs by pairs of elves.
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.

5. 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.