The COMLEX Syntax Project

Ralph Grishman and Catherine Macleod and Susanne Wolff

Department of Computer Science
New York University
New York, NY 10003

Developing more shareable resources to support natural language analysis will make it easier and cheaper to create new language processing applications and to support research in computational linguistics. One natural candidate for such a resource is a broad-coverage dictionary, since the work required to create such a dictionary is large but there is general agreement on at least some of the information to be recorded for each word. The Linguistic Data Consortium has begun an effort to create several such lexical resources, under the rubric "COMLEX" (COMmon LEXicon); one of these projects is the COMLEX Syntax Project.

The goal of the COMLEX Syntax Project is to create a moderately-broad-coverage shareable dictionary containing the syntactic features of English words, intended for automatic language analysis. We are initially aiming for a dictionary of 35,000 to 40,000 base forms, although this of course may be enlarged if the initial effort is positively received. The dictionary should include detailed syntactic specifications, particularly for subcategorization; our intent is to provide sufficient detail so that the information required by a number of major English analyzers can be automatically derived from the information we provide. As with other Linguistic Data Consortium resources, our intent is to provide a lexicon available without license constraint to all Consortium members. Finally, our goal is to provide an initial lexicon relatively quickly — within about a year, funding permitting. This implies a certain flexibility, where some of the features will probably be changed and refined as the coding is taking place.

1. Some COMLEX History

There is a long history of trying to design shareable or "polytheoretic" lexicons and interchange formats for lexicons. There has also been substantial work on adapting machine-readable versions of conventional dictionaries for automated language analysis using a number of systems. It is not our intent to review this work here, but only to indicate how our particular project — COMLEX Syntax — got started.

The initial impetus was provided by Charles Wayne, the DARPA/SISTO program manager, in discussions at a meeting held at New Mexico State University in January 1992 to inaugurate the Consortium for Lexical Research. These discussions were further developed at a session at the February, 1992 DARPA Speech and Natural Language Workshop at Arden House; a number of proposals were offered there for both interchange standards and shareable dictionaries and grammars. At a subsequent DARPA meeting in July 1992 these ideas crystallized into a proposal by James Pustejovsky and Ralph Grishman to the Linguistic Data Consortium to fund a COMLEX effort.

Starting from this general proposal, a detailed and formal specification of the syntactic features to be encoded in the lexicon was developed at New York University in the fall of 1992. These specifications were presented at several meetings, at NYU, at the Univ. of Pennsylvania, and at New Mexico State University, and form the basis for the project described here.

2. Structure of the Entries

Each entry is organized as a nested set of feature-value lists, using a Lisp-style notation. Each list consists of a type symbol followed by zero or more keyword-value pairs. Each value may in turn be an atom, a string, a list of strings, feature-value list, or a list of feature-value lists. This is similar in appearance to the typed feature structures which have been used in some other computer lexicons, although we have not yet made any significant use of the inheritance potential of these structures.

Sample dictionary entries are shown in Figure 1. The first symbol gives the part of speech; a word with several parts of speech will have several dictionary entries, one for each part of speech. Each entry has an :orth feature, giving the base form of the word. Nouns, verbs, and adjectives with irregular morphology will have features for the irregular forms :plural, :past, :pastpart, etc. Words which take complements will have a subcategorization (:subc) feature. For example, the verb "abandon" can occur with a noun phrase followed by a prepositional phrase with the preposition "to" (e.g., "I abandoned him to the linguists.") or with just a noun phrase complement ("I abandoned the ship."). Other syntactic features are recorded under :features. For example, the noun "abandon" is marked as (countable :pval ("with")), indicating that it must appear in the singular with a determiner unless

```
(verb         :orth "abandon" :subc ((np-pp :pval ("to")) (np)))
(noun         :orth "abandon" :features ((countable :pval ("with"))))
(prep         :orth "above")
(adverb       :orth "above")
(adjective    :orth "above" :features ((ainrn) (apreq)))
(verb         :orth "abstain" :subc ((intrans)
                                     (pp :pval ("from"))
                                     (p-ing-sc :pval ("from"))))
(verb         :orth "accept" :subc ((np) (that-s) (np-as-np)))
(noun         :orth "acceptance")
```

Figure 1: Sample COMLEX Syntax dictionary entries.

it is preceded by the preposition "with".

Other formats have been suggested for dictionary sharing, notably those developed under the Text Encoding Initiative using SGML (Standard Generalized Markup Language). We do not expect that it would be difficult to map the completed lexicon into one of these formats if desired. In addition, some dictionary standards require an entry for each inflected form, whereas COMLEX will have an entry for each base form (lemma). COMLEX has taken this approach in order to avoid having duplicate and possibly inconsistent information for different inflected forms (e.g., for subcategorization). It is straightforward, however, to "expand" the dictionary to have one entry for each inflected form.

In addition to the information shown, each entry will have revision control information: information on by whom and when it was created, and by whom and when it was revised. We are also intending to include frequency information, initially just at the part-of-speech level, but eventually at the subcategorization frame level as well.

3. Subcategorization

We have paid particular attention to providing detailed subcategorization information (information about complement structure), both for verbs and for those nouns and adjectives which do take complements. The names for the different complement types are based on the conventions used in the Brandeis verb lexicon, where each complement is designated by the names of its constituents, together with a few tags to indicate things such as control phenomena. Each complement type is formally defined by a frame (see Figure 2). The frame includes the constituent structure, :cs, the grammatical structure, :gs, one or more :features, and one or more examples, :ex.[1] The constituent structure lists the constituents in sequence; the grammatical structure indicates the functional role played by each constituent. The elements of the constituent structure are indexed, and these indices are referenced in the grammatical structure field (in vp-frames, the index "1" in the grammatical structures refers to the subject of the verb).

Three verb frames are shown in Figure 2. The first, s, is for full sentential complements with an optional "that" complementizer. The second and third frames both represent infinitival complements, and differ only in their functional structure. The to-inf-sc frame is for subject-control verbs — verbs for which the surface subject is the functional subject of both the matrix and embedded clauses. The notation :subject 1 in the :cs field indicates that surface subject is the subject of the embedded clause, while the :subject 1 in the :gs field indicates that it is the subject of the matrix clause. The indication :features (:control subject) provides this information redundantly; we include both indications in case one is more convenient for particular dictionary users. The to-inf-rs frame is for raising-to-subject verbs — verbs for which the surface subject is the functional subject only of the embedded clause. The functional subject position in the matrix clause is unfilled, as indicated by the notation :gs (:subject () :comp 2).

We have compared our subcategorization codes to those used by a number of other major lexicon projects in order to insure that our codes are reasonably complete and that it would not be too difficult to map our codes into those of other systems. Among the projects we have studied are the Brandeis Verb Lexicon[2], the ACQUILEX Project [3], the NYU Linguistic String Project [2], and the Oxford Advanced Learner's Dictionary [1].

[1] The general format used for constituent structures was suggested by Bob Ingria for the DARPA Common Lexicon.

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

```
(vp-frame s          :cs ((s 2 :that-comp optional))
                     :gs (:subject 1 :comp 2)
                     :ex "they thought (that) he was always late")

(vp-frame to-inf-sc :cs ((vp 2 :mood to-infinitive :subject 1))
                     :features (:control subject)
                     :gs (:subject 1 :comp 2)
                     :ex "I wanted to come.")

(vp-frame to-inf-rs :cs ((vp 2 :mood to-infinitive :subject 1))
                     :features (:raising subject)
                     :gs (:subject () :comp 2)
                     :ex "I seemed to wilt.")
```

Figure 2: Sample COMLEX Syntax subcategorization frames.

4. Creation and Verification

We are deriving the word and part-of-speech lists for COMLEX from two sources: (1) the dictionary file prepared by Prof. Roger Mitton, which was derived from the Oxford Advanced Learner's Dictionary; (2) word lists (with frequency information) obtained from corpora and tagged corpora. We are already using the "joint ventures" corpus prepared for the Tipster information extraction task (and for MUC-5); we expect to employ other and larger corpora in the future.

Using these word lists, a number of part-time staff members will manually assign syntactic features to each word. These staff members will have access to several conventional dictionaries as well as a large on-line text concordance.

We intend to use a variety of techniques to verify the dictionary information. A portion of the dictionary will be coded twice; a comparison of the resulting entries will give us some estimate of the error rate. We will compare the subcategorization information produced by our codes with the codes derived from the Oxford Advanced Learner's Dictionary, and review discrepancies.[3] For the less frequent features, we will list all the words assigned a particular feature; this often will point up inconsistencies in coders' judgements. Finally, we hope in the near future to couple the assignment of subcategorization features with the tagging of a corpus.

5. Status

As of April 1993,

- the formal specifications have been further revised and are now largely complete
- a manual has been prepared with more extensive narrative descriptions of the classes to assist coders in preparing dictionary entries
- a menu-based program has been developed for rapid preparation of dictionary entries; this program is coded in Lisp using the Garnet graphical user interface package
- an initial dictionary of all closed-class words (those with parts of speech other than noun, verb, adjective, and adverb) has been prepared

Creation of dictionary entries for the open-class words is just beginning. We hope that corpus tagging of word instances with respect to their subcategorization pattern can begin in the summer and proceed in parallel with the dictionary preparation effort.

6. Acknowledgement

This material is based upon work supported by the Advanced Research Projects Agency through the Office of Naval Research under Award No. MDA972-92-J-1016 and The Trustees of the University of Pennsylvania.

References

1. A. S. Hornby, ed. *Oxford Advanced Learner's Dictionary of Current English*, 1980.
2. Naomi Sager. *Natural Language Information Processing*, Addison-Wesley, 1981.
3. Antonio Sanfilippo. LKB Encoding of Lexical Knowledge. In *Default Inheritance in Unification-Based Approaches to the Lexicon*, T. Briscoe, A. Copestake, and V. de Pavia, eds., Cambridge University Press, 1992.

[3] We would hope to obtain permission to compare our dictionary with other broad-coverage dictionaries, and use the result to further improve our dictionary.