Catherine Macleod, Ralph Grishman & Adam Meyers,
New York University

COMLEX Syntax: An On-Line Dictionary for Natural Language Processing

Abstract

COMLEX Syntax is a large (38,000 head words) on-line syntactic dictionary of English, developed at New York University under the auspices of the Linguistic Data Consortium (LDC). It was envisioned primarily as a tool to aid in the parsing of natural language by computers. To this end, it contains an exceptionally detailed set of syntactic features and complements for the major parts of speech (nouns: 9 features and 9 complements; adjectives: 7 features and 14 complements; verbs: 5 features and 92 complements). The lexicon also contains, for 750 common verbs, references to citations (tags) in a large corpus (100 MB). This corpus is also available from the LDC. These citations can be used to gather frequency-of-occurrence statistics for the complements of these verbs and have also served as a quality check on the dictionary. COMLEX Syntax Version 2.2 (the tagged version) is now available to members of the LDC for research and commercial purposes with minimal legal encumbrances.

Introduction

COMLEX Syntax, developed by the Proteus Project at New York University, is one of the lexical resources available from the Linguistic Data Consortium (LDC). These resources are intended to serve the whole Natural Language Processing (NLP) community and to be used by researchers from both universities and commercial enterprises with minimum legal restrictions. Although there are large commercial dictionaries obtainable on-line, they were not specifically designed for NLP and may have stringent licensing restrictions on their commercial use.

In order to keep COMLEX Syntax legally unencumbered, we entered all the information from scratch.[1] Using a menu-based entering program developed at NYU, three to four linguistics graduate students have been working part-time for over two years. They consult hard copy dictionaries, an on-line concordance[2] and their own judgement as native speakers of English. The first year's efforts, reported in COLING94 (Grishman 1994), resulted in a dictionary with detailed syntactic information about adjectives, nouns and verbs. The second year, at least 100 examples of text for each of 750 frequent verbs were tagged with

COMLEX classes. The first version of COMLEX Syntax was delivered to the LDC in May, 1994 and the tagged version was delivered in August, 1995.

Background

In our desire to create a dictionary which would be useful to the whole NLP community, we endeavored to include detailed yet "theory neutral" syntactic information. As far as possible, we have used generally recognized linguistic terminology (i.e. noun, verb, adjective, preposition, adverb, and, for our complement names, their phrasal expansions into np, vp, adjp, etc.).

COMLEX Syntax's features and complements are based primarily on the classes developed by New York University's Linguistic String Project (LSP)(Fitzpatrick 1981). We selected the LSP classes because their dictionary was specifically designed for use in machine parsing of natural language, its coverage is very broad and its classes well defined. Furthermore, it included classification for nouns and adjectives as well as verbs.

We did consult several other major lexicons used for automated language analysis, to make sure that we could capture the distinctions they recognized. The Oxford Advanced Learner's Dictionary (OALD) (Horny 1980) was consulted to check our verb complement coverage against their verb codes. We checked The Longman Dictionary of Contemporary English (LDOCE) (Proctor 1978) against our adjective coverage, since LDOCE classifies both nouns and adjectives for sentential complements. As a result we modified the COMLEX classes so as to be able to recover both LDOCE and LSP sentential complements (see Table 1).

We also looked closely at the Brandeis Verb Lexicon[3] which, although not as broad coverage as the dictionaries cited above, has a very detailed analysis of verb complements and is widely known in the NLP community (in the U.S. at least). Brandeis class names are productively generated from the set of complement phrases which can follow a verb, each complement class name consisting of strings of elements separated by hyphens. COMLEX uses these strings of elements to make class membership transparent, but unlike Brandeis, assumes a fixed well-defined set of complements. In Brandeis' notation, TOVP (to infinitival), DO (direct object), OC (feature: object control) can be combined into the complement DO-TOVP-OC; this corresponds to the COMLEX complement NP-TO-INF-OC. Having a fixed set of complements allows us to explicitly define the significance of each complement name, as will be shown below. In particular, it enables us to associate one complement name with several structures, thus allowing us to capture entailment and co-occurrence relations among certain complements. It also makes it easier to create menus in our entry program for defining new words, and to specify mappings between COMLEX classes and the classes in other dictionaries.

COMLEX also adopts a version of Brandeis' notion of raising and control phenomena, which while not strictly theory neutral, allows the reconstruction of the predicate argument structure where there is a missing subject in infinitival and gerundial expressions. For example, by adding object control (-OC) to an infinitival complement (NP-TO-INF-OC) we provide the information that the matrix sentence object is both the logical object of the matrix verb and the logical subject of the infinitive. *Persuade* is a verb that takes that type of complement. Therefore we know that in the sentence, *Robert persuaded his son to go Robert* persuades *his son* and *his son* goes. Likewise, for the subject control (-SC) complements, we can predict that the matrix subject is the logical subject of both matrix and subordinate clauses. *Promise* has this complement (NP-TO-INF-SC). In the sentence, *Jason promised his son to go*, *Jason* is both promising and going, in other words *Jason* is the logical subject of both verbs.

Lastly, we studied the English verb classification developed by Sanfilippo for the ACQUILEX project (Sanfilippo 1992) and found that basically COMLEX classes covered his syntactic classifications.

Table 2 shows some mappings of COMLEX verb complement structures to LSP, OALD, Brandeis and the ACQUILEX codes. A direct mapping from COMLEX to OALD, LSP, and Brandeis is straightforward, but since the ACQUILEX notation is very complex, containing semantic information as well, COMLEX entries would need to be augmented to cover this type of information.

The entering of the dictionary was accomplished during the first year of the project. The second year (1994–1995) was spent tagging 100 citations for each of 750 most frequent verbs. Each tag indicates the COMLEX class for one instance of that verb, as it occurs in a citation from our 100 MB corpus. Tagged entries include pointers to examples of usage, similar to those found in most hard copy dictionaries. These tags are intended primarily to provide frequency statistics for computational projects. However, there are other possible uses. In one previous study, these tagged entries were used to make a connection with WordNet (Miller 1990) (a large on-line thesaurus-like facility) in an attempt to relate the WordNet semantic classes (synsets) of a verb to the verb's complements. This is described in a paper presented at the 1994 International Workshop on Directions of Lexical Research in Beijing, China (Macleod 1994).

The Structure of the Dictionary

The COMLEX entry is written in a Lisp-like[4] notation, consisting of the part of speech (noun, verb, adverb, preposition .....), the word itself (introduced by the key word :orth), a list of the features (:features), a list of the complements or subcategorizations (:subc) and a list of tags (:tags), as appropriate. Each tag lists the byte number in the corpus where the word appears, the original source of that example and the COMLEX class (:label). A sample of dictionary entries is found in Figure 1.

As shown in the sample, punctuation is defined as a WORD and marked :POS *NONE* indicating that it has no regular part of speech (POS). Words other than nouns, adjectives and verbs are entered with their part of speech but have no further syntactic classes. Items with irregular morphology (*feet* not *foots*, *came* not *comed*, *better* not *gooder*) are specifically entered, but regular morphology is described in the COMLEX Syntax Reference Manual (Macleod).[5] If two words have the same orthography but are different parts of speech, they are given a separate entry for each POS (*foot* as noun and *foot* as verb). However, we do not sense differentiate within the same part of speech. Therefore the entry for *bass* is classified as human (feature NHUMAN) for the singer, even though the same orthography encompasses the instrument and the fish which are, of course, not human. If any sense of a word has a particular feature or complement, it is assigned to the word. Thus a word could potentially have conflicting features.

Complements

Features and complements are defined in the reference manual (Macleod). Complements are represented by frames and frame groups (See Figure 2). Frames specify the surface structure, the constituent structure, and the control/raising features (if any) and give an example. A complement name defined by a frame group represents a set of frames. Frame groups capture relationships of complement alternation (np-for-np) or co-occurrence (wh-s) and save space, since the frame-group name represents more than one complement.

In Figure 2, the frame-group NP-FOR-NP represents the benefactive alternation (Levin 1993) with the frames *NP-FOR-NP and *NP-NP, this group also includes the heavy np shift *FOR-NP-NP.[6] The frame NP-TO-INF-OC illustrates a complement with the object control feature. The WH-S frame-group includes finite clauses introduced by *whether, if* and *what* (the latter contains an omitted noun phrase) and infinitival clauses introduced by the same complementizers (*wh-to-inf and *what-to-inf) which have been eliminated from the figure in the interest of space.

Features

Though features for nouns, adjectives and verbs are defined syntactically they are sometimes semantic in nature as well.[7]

In the sample dictionary (Figure 2), *foot* has two features: NUNIT and COUNTABLE. The definition for NUNIT is as follows: "a noun is an NUNIT if it can occur as the noun in the measure sequence quantifier-noun followed by pp or adj of dimension", e.g. *two INCHES in length/ two INCHES long* (Macleod). The countable feature determines whether an article is required for this noun in the singular. A noun phrase containing an instance of this singular noun as the head noun is ill-formed unless it also contains a determiner. This feature allows two exceptions :PVAL and :PREDNOUN; if the entry has one of these key words, it may occur without a determiner in those environments (as the object of the specified preposition or as a predicate complement). For example, *her foot is small/ *foot is small* but *he traveled on foot* is fine because the entry includes :pval 'on'.

The adjective feature GRADABLE by itself indicates that the comparative and superlative forms of the adjective are formed using *more* and *most* plus the adjective; an adjective is marked GRADABLE :ER-EST if the comparative and superlative forms are derived morphologically (*happy, happier, happiest*) or GRADABLE :BOTH if it occurs both ways (*more happy, most happy*). The verb feature found in the sample dictionary is VSAY. This denotes a verb that can occur with a quoted statement (see the tagged example in the next subsection).

It should be noted that our feature definitions are not exhaustive environments but only supply the condition which must be met in order for a word 'to be identified as having a particular feature. For example, although the definition of the verb feature VSAY requires that this type of verb be able to occur with a quoted statement, it is not the only environment where it occurs. Also we would like to note that this dictionary in not meant to be able to handle literary usages where words are coerced into different classes. There are many examples of this. Examples like *a grief ago* or *"Not me sir", she blushed* will not result in our adding *grief* as a time noun (NTIME) nor categorizing *blush* as a VSAY in COMLEX. These classes are intended to capture ordinary usage.

Tags

The tags shown in Figure 1 are just the first three tags entered, since space considerations prevent us from listing all 100 tags here. The notation for the tags shows the :BYTE-NUMBER, which is the starting byte-number of the verb instance, :SOURCE, which here is the Brown Corpus, and :LABEL (the name of the complement or feature which occurs in the tagged sentence). The sentence corresponding to the VSAY citation (with the tagged word in upper case) is *Spahnie doesn't know how to merely go through the motions REMARKED Enos Slaughter, another all-out guy, who played rightfield that day and popped one over the clubhouse.*

The tagging is described in COMLEX Syntax 2.0 Manual for Tagged Entries (Meyers) which is available from the LDC or from New York University (by ftp).

Summary

COMLEX Syntax was specifically designed and hand-entered as a tool for computational applications. Therefore, we believe that it is both richer and more consistent in coding than the machine-readable versions of dictionaries which were intended primarily for publication and for general use or pedagogical purposes. Our own usage of COMLEX and the feedback we have gotten from other users seems to validate this viewpoint.

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.

Notes

1. We would like to thank the Oxford University Press for permission to use part-of-speech lists from the OALD for our dictionary.
2. The corpus is comprised of 100 MB of text including the complete Brown Corpus, newspaper articles from the San Jose Mercury and the Wall Street Journal, Department of Energy Abstracts and some literature from the Library of Congress.
3. Developed by J. Grimshaw and R. Jackendoff under grant NSF IST--81--2040.
4. We are considering creating a mapping of our notation to SGML because that format may be preferred for some computational studies in the humanities outside of computational linguistics.
5. Available by anonymous ftp from cs.nyu.edu in the directory pub/html/comlex.html or from our web site http://cs.nyu.edu/cs/faculty/grishman/comlex.html .
6. The complement names preceded by an asterisk do not appear in the dictionary but are used in the definition of the frame-group.
7. The feature nhuman which seems to be semantic, is defined as follows: A noun with the feature NHUMAN can occur as the head noun of relative clauses introduced by *who* or *whom*. For example: *The man who is in the room* not **The book who is in the room.*

Tables

COMLEX	LSP	LDOCE	Example Sentence
extrap-adj-that-s	asent1: athats	F5	it is curious *that he left*
that-s-adj	asent3: athats	F5	they were aware *that he was sick*
extrap-adj-s	asent1: athats	F5a	it is probable *(that) they left*
s-adj	asent3: athats	F5a	he was sure *(that) she knew*

Table 1: Mapping of Adjective Complements from COMLEX to LSP and LDOCE

COMLEX	LSP	OALD	ACQUILEX (Sanfilippo)	Brandeis	Example
np-to-np	npn pval: `to' nn	vp12a vp13a	obl-trans-sign ditrans-sign	do-tonp io-do	he gave *the book to her* he gave *her the book*
extrap-np-s	vsent1		extrap-comp-trans-sign extrap-equi-trans-vpinf-sign		it pleases them that she went it pleases them to go
seem-s seem-to-np-s to-inf-rs	vsent4 vsent4 tovo	vp4e	extrap-comp-intrans-sign extrap-obl-comp-intrans-Sfin-sign subj-raising-intrans-vpinf-sign	tovp-sc-rs	it seems (that) they left it seems to her that he was wrong he seems to sleep

Table 2: Mapping of Verb Subcategorizations

```
(WORD:ORTH “;“ :POS *NONE*)
(SCONJ:ORTH “after“)
(NOUN:ORTH “bass“:FEATURES ((NHUMAN))
(ADJECTIVE:ORTH “calm“:FEATURES ((GRADABLE :BOTH T)))
(NOUN:ORTH “foot“:PLURAL “feet“
            :FEATURES ((NUNIT)(COUNTABLE :PVAL (“on“))))
(VERB :ORTH “foot“ :SUBC ((NP)))
(PREP :ORTH “for“)
(PRONOUN:ORTH “he“ :FEATURES ((SINGULAR)(NOMINATIVE)))
(ADVERB:ORTH “honestly“)
(VERB :ORTH “remark“:SUBC((PP-THAT-S :PVAL (“to“)) (S)
                   (PP :PVAL (“on“ “about“)) (THAT-S) (HOW-S)
            :FEATURES ((VSAY))
            :TAGS ((TAG :BYTE-NUMBER 6860726
                        :SOURCE “brown“
                        :LABEL (S))
                   (TAG :BYTE-NUMBER 6854938
                        :SOURCE “brown“
                        :LABEL (THAT-S))
                   (TAG :BYTE-NUMBER 6776443
                        :SOURCE “brown“
                        :LABEL (VSAY))))
(NOUN:ORTH “remark“ :subc ((NOUN-THAT-S)(NOUN-BE-THAT-S)))
```

Figure 1: Sample COMLEX Syntax dictionary entries.

```
(frame-group np-for-np(*np-np *np-for-np *for-np-np))
(vp-frame *np-for-np:cs ((np 2) “for“ (np 3))
      :gs (:subject 1, :obj 2, :obj2 3)
      :ex “she bought a book for him.“)
(vp-frame *for-np-np:cs (“for“ (np 2) (np 3))
      :gs (:subject 1, :obj 3, :obj2 2)
      :ex “she bought for him a book that she had found to be in-
teresting.“)
(vp-frame *np-np:cs ((np 2) (np 3))
      :gs (:subject 1, :obj 3, :obj2 2)
      :ex “she bought him a book.“
(vp-frame np-to-inf-oc:cs ((np 2)(vp 3 :mood to-infinitive :subject
2))
      :features (:control object)
      :gs (:subject 1, :obj 2, :comp 3)
      :ex “I advised Mary to go.“)
(frame-group wh-s(*wh-s *wh-to-inf *what-s *what-to-inf))
(vp-frame *wh-s:cs (s 2 :q (if wheth))
      :gs (:subject 1, :comp 2)
      :ex “he asked whether he should come.“)
(vp-frame *what-s:cs (s 2 :q (what 3) :omission 3)
      :gs (:subject 1, :comp 2)
      :ex “he asked what he should do.“)
```

Figure 2: Sample COMLEX Syntax verb subcategorization frames.

References

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

Grishman, Ralph, Catherine Macleod and Adam Meyers 1994. Comlex Syntax: Building a Computational Lexicon, in: *The Proceedings of COLING94.*

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

Levin, Beth 1993. *English Verb Classes and Alternations.* The University of Chicago Press, pp 48–49.

Macleod, Catherine and Ralph Grishman. COMLEX Syntax Reference Manual. Proteus Project, Computer Science Department, New York University.

Macleod, Catherine, Ralph Grishman, and Adam Meyers 1994. Developing Multiply Tagged Corpora for Lexical Research, in: *The Proceedings of the Post-Coling International Workshop on Directions of Lexical Research, Beijing, China, August, 1994*, pp 11–22.

Meyers, Adam, Catherine Macleod, and Ralph Grishman. COMLEX Syntax 2.0 Manual for Tagged Entries. Proteus Project, Computer Science Department, New York University.

Miller, George (ed.) 1990. WordNet: An on-line lexical database, in: *International Journal of Lexicography* (special issue), 3(4), pp 235–312.

Proctor, P., (ed.) 1978. *Longman Dictionary of Contemporary English.* Longman.

Sanfilippo, Antonio 1992. LKB encoding of lexical knowledge. in: T. Briscoe, A. Copestake, and V. de Pavi (eds), *Default Inheritance in Unification-Based Approaches to the Lexicon.* Cambridge University Press.