Carol Peters
*Istituto di Elaborazione dell'Informazione, CNR, Pisa*

Stefano Federici
*Istituto di Linguistica Computazionale, CNR, Pisa*

Simonetta Montemagni
*Istituto di Linguistica Computazionale, CNR, Pisa*

Nicoletta Calzolari
*Istituto di Linguistica Computazionale, CNR, Pisa*

# From Machine Readable Dictionaries to Lexicons for NLP: the Cobuild Dictionaries – a Different Approach

## Abstract

We describe the results of a syntactic–semantic parser for Cobuild dictionary definitions. Unlike previous work on the automatic analysis of machine readable dictionaries, the particular structure of the Cobuild definition allows us to derive information that classifies the lexical item mainly in terms of the selectional restrictions or preferences encoded on its arguments. The resulting formalized lexical entries contain data that has generally been lacking in other lexical representations but which is expected to be very useful in a wide range of NLP purposes. We show how this information can be used in dictionary sense disambiguation by creating links throughout the lexicon both on the paradigmatic and the syntagmatic axes.

## 1. Introduction

It is widely recognized that machine readable dictionaries are important sources of lexical data which, suitably analysed, extracted and formalized, could be used effectively in the construction of lexical components for NLP applications. Recent studies in this direction have concentrated on classifying the headword in terms of the particular semantic features that can be derived from the genus items and their differentiae, and representing the lexicon as a hierarchical inheritance network (see, for example, the ACQUILEX project as described in Calzolari et al., forthcoming). In the paper we describe a different strategy that has been adopted to acquire lexical information from the Cobuild Student's dictionary. This approach exploits two special features of the Cobuild range of dictionaries: (i) they are compiled on the basis of the evidence provided by a very large corpus of contemporary spoken and written English; (ii) the definitions appear as complete natural language sentences, i.e. with the definiendum inserted in its typical sentential context. Working on this data, our objective has thus been

to study a method to represent formally the actual usage of words. The aim is to define a route from the actual dictionary text to formal grammar, representing the distinctive patterns of language in use in a language–independent formalism. The work described is part of the CEC project ET–10/51, "Semantic Analysis, using a Natural Language Dictionary", carried out in collaboration with the Universities of Birmingham (coordinator) and Bochum.

## 2. The Cobuild defining strategy

The Cobuild dictionaries have been constructed on the assumption that words have sense only in context. The analysis of the corpus evidenced that a given lexical item very often reveals not just a typical syntax but also typical patterns of lexical co–occurrence, and that particular structural and/or lexical patterns are frequently associated with particular senses (Sinclair 1987). Thus, while in traditional lexicography statements are made about what words mean but not much is shown on their use, the innovative form of the Cobuild definition not only explains the meaning of the headword (in the right hand side – RHS) but also illustrates its use by presenting it in its typical syntactic and lexical/semantic context (the left hand side – LHS).

The intention is not just to help the human user in decoding a text but to provide useful models for encoding. Whenever possible, typical grammar structures and typical cooccurring items are given for each sense of a headword. For example, the verb **diagnose** is defined in the Student's dictionary by "When a doctor **diagnoses** a disease that someone has, he or she identifies what is wrong" . From the LHS of this definition it can be inferred that the required or preferred subject of this verb is a doctor, the required direct object a disease, and that this disease (i.e. a disease diagnosed by a doctor) is particular of human beings; in addition to stating the meaning, the RHS indirectly assigns the features human, male/female to the argument **doctor** and inanimate to **disease**. This kind of information is not found in the same way in other dictionaries, cfr. OALDCE: "determine the nature of (esp a disease) from observation of symptoms", from which we can only infer that **disease** is related in some way with **diagnose** but the exact nature of the relationship is not immediately recognizable from the definition text. Indeed different senses of a word can often be distinguished by the different kinds of arguments or collocates associated with them. For example, the two senses of **adore** in the *Student's dictionary* are explained as follows: 1. If you **adore** someone, you love and admire them; 2. If you **adore** something, you like it very much. These senses are differentiated by the fact that while in both cases the typical subject is human, in the first the required object is also human whereas in the second it is inanimate. Other important information on the user perspective of the verb, e.g. whether socially reprehensible, possible, inherent, is also intentionally implied in the definitions. This is given in the examples above by the use of 'when' or 'if' as initial operators and the choice

of 'you' or 'someone' as indicators of human arguments. For a discussion of the significance of the Cobuild defining strategies, see Hanks (1987).

The LHS of the Cobuild definition (integrated by certain data from the RHS) thus contains information that cannot be found systematically in traditional dictionaries; in fact, such dictionaries provide this kind of information only occasionally, in the form of example sentences and very rarely as specifications within the definition text. By contrast, in Cobuild this information is usually encoded in a consistent, coherent way. It is the linking of a number of elements, i.e. (i) the statement of meaning, (ii) the syntactic environment, (iii) the selectional preferences or restrictions on arguments, and (iv) information on the user perspective of the verb, which, on the one hand, is unique to Cobuild and, on the other, is of primary importance for the solution of many problems in NLP. For this reason we have focussed our efforts on extracting and representing this kind of information in a re–usable formal framework.

### 3. Analysing and representing the lexical entries

In this section we describe the methodology we followed in order to extract syntactic, semantic and also pragmatic information from the Cobuild definitions and to formalize it in a lexical representation language. The first task was to study in depth samples of the syntactically 'chunked' definition data supplied by our partners at Birmingham University so that we could identify the different types of information contained, and the ways in which they had been represented in the dictionary. As has been stated, we were particularly interested in the information on syntactic and lexical/semantic constraints on and preferences of the arguments of a lexical item that could be derived from the LHS. From our first analysis of the data, it became clear that the very regular structure and defining formulae employed by Cobuild to encode this kind of implicit data could be exploited in order to extract it. We next had to decide on the best way to formalize and represent the lexical entries that we were constructing in a computationally tractable and re–usable way. Our objective has been to produce results that would be viable and exploitable both by the human user and the machine. Thus we adopted a two stage approach. In the first stage all the information extracted was mapped onto an Intermediate Template (IT); in a second stage the different types of information extracted were evaluated with respect to their representability and utility for NLP and then converted into a Typed Feature Structure (TFS) formalism. Whereas the IT has been conceived mainly as a theory–neutral and re–usable representation format, useful for both human users and the machine, the TFS representation format has mainly been chosen for its computational tractability, and more specifically for its integrability in NLP components. Full details on the specialised parser that was designed and developed for this purpose can be found in ET–10/51 Group (1993).

For each entry, the IT contains tagged, detailed and explicit orthographic, phonetic, morpho–syntactic, syntactic and semantic information. Thus, the IT presents explicitly much information which is only contained implicitly in the printed dictionary. An example of the format of our results at this stage are given in Figure 1 for the definition of **apply 4**: "If you apply a rule, system, or skill, you use it in a situation or activity"

```
sense_no          : 4
lemma             : apply
entry_info        : entry                : apply
genus_info        : prov_superordinate1  : use
                    isa                  : use
                    genus_prep           : in
inflection        : apply applies applying applied
gram              : VB with OBJ
voice             : active
inference         : possible likely
subj_info         : subj_features1       : human
obj_info          : obj1                 : specific: rule
                    obj_features1        : inanimate,+count
                    obj2                 : specific: system
                    obj_features2        : inanimate,+count
                    obj3                 : specific: skill
                    obj_features3        : inanimate,+count
usage_info        : formality            : normal
                    style                : normal
```

Figure 1: **apply** 4 represented on the Intermediate Template

As already stated, the 'inference' attribute classifies the verb in terms of the action perspective. The value here is derived from 'If you' in the definition. 'Subj_info' and 'obj_info' are complex attributes which formalize the implicit selectional restrictions, i.e. that **apply** in this sense prefers a human subject and inanimate objects typically exemplified by rule, system and skill. The feature 'human' associated with the subject is inferred by the pronoun 'you' occupying equivalent positions in both the LHS and the RHS. The features 'inanimate' and '+count' associated with the object have been derived by the use of the indefinite article in the LHS and the presence of a so–called matching pronoun 'it' in the RHS. Data of this kind on preferred arguments and their semantic features are crucial for NLP applications, but typically difficult to derive from other sources and expensive to encode extensively and consistently by hand. The genus_info complex attribute contains both the provisional genus term provided by the Birmingham data (tagged either as superordinate or synonym) and the results of our analysis of this data. Our parser examines this data in order to derive significant values depending on the semantic relation between the genus term and the headword, e.g. synonymy, hyperonymy, set of, part of, member of, etc.

Our TFS entry is mainly based on the theoretical notions of the HPSG grammar (Pollard and Sag 1987, forthcoming), which we have enriched and adapted to include and represent the information we extract from the Cobuild dictionary. The reason behind this choice is two–fold. First, unlike other formalisms such as DATR, HPSG has the advantage of being not specific to lexical representation; this makes it much easier to integrate our lexical entries in NLP systems based on HPSG, for testing and use. Secondly, HPSG has been designed as an integrated theory of natural language syntax and semantics; this makes it easier to formalize one of the main assumptions behind the Cobuild dictionary, i.e. the interlocked dependency of syntactic, lexical and semantic properties in the definition of lexical items. In the following we limit the discussion to showing how our TFS entry differs with respect to HPSG. The semantic preferences imposed by the verb on its arguments have been encoded as specifications on the elements of the subcategorization list, here jointly represented by the SUBJ and COMPS attributes. As can be noted, the 'inanimate' feature no longer appears as a restriction on the possible objects of apply 4; this follows from the fact that here 'rule', 'system' and 'skill' refer to types which are part of a semantic ontology, and their 'inanimateness' is implied by their definition as subtypes of a more general type 'inanimate'. Similar observations hold for the LEXSEM attribute, encoding the information extracted from the RHS of the definition, and particularly the genus information. Thus the TFS entry really contains much more data than is first evident, represented by the information contained in the underlying type system. In Figure 2, we see how information that had been extracted from the definition statement for **apply** 4 and mapped onto the IT has been automatically converted into the TFS format.

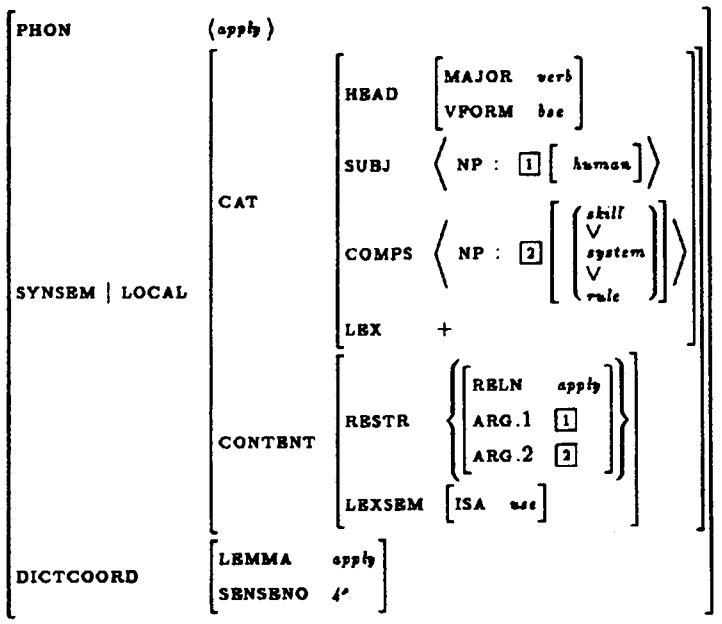


Figure 2: **apply** 4 represented in our TFS formalism

By comparing the IT and the TFS entry, we can see how they meet the needs of different kinds of users. Our Intermediate Template has been designed as the most convenient structure for a first computational model of the lexical information derived from the Cobuild definitions. User friendly interfaces could be easily implemented to make the information contained in it readily available for different kinds of human users requiring detailed information on a lexical item, its syntactic and semantic properties and its usage, e.g. translators, language learners, etc. The user would have direct dynamic access to all the information contained in the lexical entries, wherever it has been stored, and in an interpreted form, instead of being bound by the restrictions imposed by the static alphabetical ordering of the printed volume. On the other hand, we feel that the main use for the TFS entries will be in the construction of computational lexicons and in machine–driven NLP applications, e.g. in systems for analysis and generation, in word–sense disambiguation, etc.

An important application for both formats could be in computer assisted lexicography. The IT should be useful for the lexicographer employed in day by day dictionary compilation. In fact, although the current trend in computational linguistics is that of generating TFS lexicons, this kind of

representation is often not too popular with the lexicographer who finds the requirements of the formalism tend to make it 'heavy' to handle and not immediately interpretable as much of the information is underlying, i.e. contained within the type system. The IT also has the advantage that it can represent information so far not handled by standard TFS formalisms. The TFS entries, instead, could be used by the overall dictionary editor in a revision of the source or in the production of a new dictionary. The representation of the lexical entries in terms of a TFS enforces a coherent structuring of the type system behind the lexical representations, and thus encourages coherence in design of new dictionaries and assists the correction of inconsistencies in the original, by evidencing clearly what is really pertinent in the definition of different kinds of entries.

## 4. A strategy for sense disambiguation in the dictionary

Once our parser has been applied throughout the whole dictionary, over wide classes of definitions, we hope to implement a procedure now being experimented which exploits the syntactic–semantic information extracted for each lexical item to create, where possible, disambiguated, direct links on both (i) the paradigmatic axis and (ii) the syntagmatic axis, i.e. in the first case between the item and the correct sense of the relevant dictionary entry for the genus term, in the second case between the lexical item and the correct sense of its different arguments and modifiers.

### 4.1 Genus term disambiguation

The paradigmatic links such as synonymy, hyponymy, hyperonymy, as well as meronymy, are those usually extracted and formalized from MRD definitions and exploited in the construction of a hierarchical lexicon (as in Acquilex and other similar projects – see, for example, Calzolari et al., forthcoming). A major problem has always been to connect the entry item to the correct sense of the genus. In their description of the Cobuild definition statements, Allport et al. (1993) point out that the two parts of the definition (the LHS and the RHS) are in a relationship of equivalence: "The left part sets up matches which the right part must match or otherwise take account of". We have seen that features that have been extracted as holding for the word being defined are shared by the genus term and can thus be used to disambiguate it, i.e. the combination of the information that has been extracted from the LHS and the RHS of one definition can then be projected onto the LHS's or RHS's of all the definitions listed by the dictionary for the lexical item corresponding to the genus term, searching for matches that will allow us to identify the right sense. For instance, if we refer back to the example of **apply 4** in the previous section: "If you **apply** a rule, system, or skill, you use it in a situation or activity", we can assign the subject (+hum) and the object (–anim, +count) preference features attached to **apply** in this

sense to the potential arguments of its superordinate **use** in order to attempt to identify which sense of **use** is implied here.

The Student's dictionary entry for **use** has 7 different senses and 3 of them are for verbs:

1. VB with OBJ If you **use** a particular thing, .....
2. VB with OBJ If you **use** a particular word or expression, .....
3. VB with OBJ If you **use** people, .....

In each case, the subj_features (+hum) which we could derive from these definitions would match with those that we have assigned to **use** as superordinate of **apply 4**. However, the best match as far as the obj_features are concerned would be with **use 1**. Sense 3 is immediately excluded as the obj_features that we derive from people include +hum, whereas in sense 2, while the subj_features would match, the values for obj:specific would clash (rule, system or skill vs. word or expression).

Another example of this kind of genus disambiguation strategy is shown by **function**, 2, VB: "If a machine or system **functions**, it works". In this case, **work** is recognised as the genus term, and our parser tags it as a synonym. The semantic features attached to **function**, and thus assigned to **work**, are represented on the IT as follows:

| | | | |
|---|---|---|---|
| subj_info | : | subj1 | : specific : machine |
| | | subj_features1 | : –anim, +count |
| | | subj2 | : specific : system |
| | | subj_features2 | : –anim, +count |

If we look at the dictionary entry for **work**, we find 16 sense divisions; 9 of these are for verbs. However, it will be seen that the best match for **work** with the above features is clearly sense no. 8: VB "If a machine or piece of equipment works, ....", where we again find machine as subject.

This strategy appears to work well in disambiguating the genus term for verbs, although at times it may only help by reducing the possibilities rather than identifying a unique sense. Other values, in addition to argument selection features, that could be used to find the best sense match are those for the grammar and the inference attributes.

We are now evaluating to what extent it is feasible to use a similar strategy to disambiguate the genus term for nouns, again using the equivalences established between the two sides of the definition to assign the features that have been attached to the entry item also to the genus term. However, our first results are less encouraging than when working with verbs. The main problem is that the Cobuild definitions for nouns tend to be less generous with collocational information on the LHS. Thus, it is not generally likely that this information will be available or sufficient in itself to permit us to disambiguate the genus term.

Let us consider a few examples. The definition for **power** 4, UNCOUNT N with SUPP, in the dictionary is: "The **power** of something is its physical strength" where, in our input data, **strength** is tagged as superordinate of **power** and 'of something' on the LHS is matched with 'its' on the RHS. Our parser gives the following output on the IT for this sense of **power**:

| | | |
|---|---|---|
| sense_no | : 4 | |
| gram | : UNCOUNT N with SUPP | |
| entry_info | : entry | : Power |
| genus_info | : prov_superordinate | : strength |
| | is–a | : strength |
| colloc_info | : colloc_prep | : preferred : of |
| | colloc_features | : –anim |

The entry for **strength** in the dictionary gives 8 definitions, all refer to nouns, but the best match with the preferred collocational information that we have transferred to **strength** in this definition on the basis of its relationship with **power** is with sense 5 UNCOUNT N "The strength of an object is its ability to withstand rough treatment or heavy weights". Cfr. 1. "Your strength is ...", 2."Strength is also courage or determination", 4. "Your strengths are ..." 6. "The strength of a feeling or opinion is ...", 7. "The strength of an opinion, argument or story is ...", 8. "The strength of a relationship is ...". The only other sense for which a match could be recognized is 3, UNCOUNT N. "You can also refer to power or influence as strength ..." where the potential matching would be between the headword of the definition we are examining (**power**) and the genus term. However, **power** in this definition seems to refer to sense 1 of the dictionary and not to sense 4. In fact, if our procedure were extended to take into account the example given with this sense in the dictionary "the strength ... of the unions" it would perhaps exclude it as a possible match.

This particular example gives a good result using the proposed method but, in general, it appears that the strategy needed for sense disambiguation is more complex for nouns than for verbs and that other kinds of information normally play a role. Consider the following definition for **function 1**, COUNT N, The **function** of something or someone is its purpose or role, where 'purpose' and 'role' are tagged as synonyms of **function** by our parser. The entry for purpose gives 3 senses:

1. COUNT N The **purpose** of something is the reason ...
2. COUNT N Your **purpose** is the thing that ....
3. UNCOUNT N **Purpose** is the feeling ...

where we find that the information we extract on collocational preferences for **function** would allow us to match its genus **purpose** against both dictionary senses 1 and 2 (Your purpose = the purpose of someone), and the

grammar information also matches in these two senses. Thus our range of choice is reduced but not exhausted.

Whereas, if we look at the following dictionary definitions for **role**, we find that the genus term for sense 1 actually includes the word 'function' (a typical example of (semi-)circularity in the definitions which lexicographers try to avoid although at times it appears to be inevitable) and it is this that will permit our procedure to select the correct sense, rather than the collocational information.

1. COUNT N with SUPP Your **role** is your position and function .... (genus: position and function)
2. COUNT N A **role** is one of the characters that .... (genus: characters)

### 4.2 Argument disambiguation

We have also examined methods to create disambiguated syntagmatic links, i.e. to be able to recognize the correct sense of the words used as arguments of the headword in the definitions. This disambiguation will be driven by both the syntactic and semantic features automatically extracted from the definitions. For instance, the Cobuild definition for **pick up 4** gives us "If you **pick up** a skill or idea ...". From this, we have extracted **skill** as one of the preferred arguments for **pick up** in this sense and, on the basis of the indefinite article, we have assigned the feature '+count' to **skill**. The dictionary entry for **skill** gives two senses: count and uncount. We can thus automatically select the right sense of **skill** when it is an argument of **pick up**. Again, using this strategy, it is not always possible to immediately identify the correct sense of an item, often we can just reduce the possibility of choice.

In any case, we think that a second strategy based on the semantic features that our analysis attributes to a given argument is more interesting. For example, with **abdicate 1**, "If a king or queen **abdicates**, he or she resigns" we identify as specific subjects 'king' and 'queen' for which, on the basis of correspondences between LHS and RHS, the features +hum, +male, and +hum, +female, respectively, have been inferred. This will permit us to map directly to the first sense of the dictionary entry for **king** which has 'man' as superordinate (in the other senses of **king**, the superordinates are (chess) piece, and playing card), and also to the first sense of **queen** with superordinate 'woman' (the other senses have (chess) piece, playing card, and also bee as their superordinates).

## 5. Towards the lexicon as a set of relations

In this way, we move from the dictionary considered as a list of distinct entries towards the construction of a series of combinatorial links, so that lexical items no longer exist in isolation but are continuously associated with all the others that can enter into relationship with them (see Calzolari (1990)

for a discussion on the lexicon as a complex set of relations). Two aspects must be stressed: (i) the dictionary itself provides the means to construct automatically disambiguated links, both in the paradigmatic and syntagmatic directions; (ii) the extraction of syntagmatic or phrasal links is a peculiar feature of this project and allows us to encode in the formal lexical entries that we are creating both their syntactic environment and the semantic preferences on their neighbours (whether arguments, modifiers, governors, etc.). From a theoretical perspective, these results, which blurr the familiar decoupling of lexis and syntax, could also be seen as an attempt to formalize the linguistic hypotheses behind the Cobuild dictionaries.

**References**

Allport, G., Barnbrook, G. and Sinclair, J. 1993. "A Grammar of Cobuild Definition Statements". Working Paper for ET–10/51. Birmingham University.

Calzolari, N. 1990. "Structure and Access in an Automated Lexicon and Related Issues". *Linguistica Computazionale* 6: 139–161.

Calzolari, N., Hagman, J., Marinai, E., Montemagni, S., Spanu, A. and Zampolli, A. (forthcoming). "Encoding Lexicographic Definitions as Typed Feature Structures" to appear in F. Beckmann, G. Heyeder (eds.), *Theorie und Praxis des Lexikons*. Berlin & New York: Walter de Gruyter.

Collins Cobuild 1990. *Student's Dictionary. London and Glasgow*: Collins.

ET–10/51 Group 1993. "The Parsing of Cobuild Definitions and Mapping of the Output to Typed Feature Structures and Bochum Logical Form", ET–10/51 Deliverable 4. Luxembourg: CEC.

Hanks, P. 1987. "Definitions and Explanations" in J.M. Sinclair (ed.), *Looking Up: An Account of the COBUILD Project in Lexical Computing*, 116–136. Birmingham: Collins COBUILD.

Hornby, A.S. 1974. *Oxford Advanced Learner's Dictionary of Current English*. Oxford: Oxford University Press.

Pollard, C., Sag, I. 1987. *Information–based syntax and semantics – Vol. 1*. Stanford – CLSL: University of Chicago Press.

Sinclair J. (1987). "Grammar in the Dictionary", in J.M. Sinclair (ed.), *Looking Up: An Account of the COBUILD Project in Lexical Computing*, 105–116. Birmingham: Collins COBUILD.