# Incorporation of Case Frames for Multi-Word Verbs Into a Lexical Database

Phyllis M. Kowalke

Elmhurst College
190 Prospect Avenue
Elmhurst, Illinois 60126
phyllisk@elmhurst.edu

**Abstract**

This paper describes the Multi-Word Case Frame (MWCF) database designed to be a component of the Illinois Institute of Technology lexical database (IITLEX). The MWCF database is a tool that will be used by future word-disambiguation research projects to determine whether a verb plus particle combination is a multi-word verb or simply a free combination of verb plus adverb or preposition. The database also contains the links necessary to trace the multi-word verb to the information in the dictionary databases.

## Introduction

The Multi-Word Case Frame (MWCF) database is one of many components of the Illinois Institute of Technology lexical (IITLEX) database (Evens et al. 1991). The MWCF database contains case frame data for multi-word verbs. The multi-word verbs in MWCF are either phrasal verbs that consist of a verb plus adverbial particle or prepositional verbs that consist of a verb plus prepositional particle.

The MWCF database is a tool to be used by future research projects. The files in the MWCF database are designed to be used for word sense disambiguation and for determining if a verb plus particle combination is a free combination or a multi-word verb. In a free combination, the verb and particle each have their own meaning while the meaning of a multi-word verb cannot be deduced from the individual components. Rules to make the above determination depend upon whether the multi-word verb is a phrasal verb or a prepositional verb.

To maximize utility of the MWCF database during word-sense disambiguation, links are included throughout the database to trace data from the verb form through the case frames back to the original dictionary databases.

Figure 1 shows the structure of the MWCF database. The database was created as a Microsoft Access relational database. Portions of the database are available as flat files via ftp. These flat files can then be imported into relational database packages.

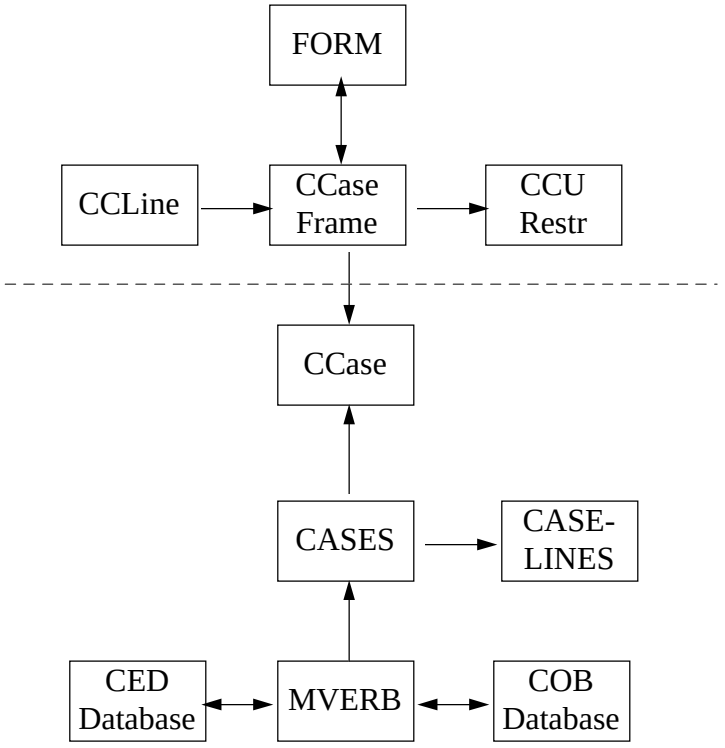


Figure 1: The MWCF Database

## Multi-Word Verbs

In grammar courses we learned that you should never end a sentence with a preposition. Following this rule, a sentence such as *Which hill did Jack run up?* is grammatically incorrect. We can correct the sentence by switching the words around to *Up which hill did Jack run?* producing a more formal sentence that follows the rule. But, what about a sentence such as *Which account did they run* up?

Switching the words to **Up which account did they run?* produces a grammatically incorrect sentence because *run up* is a two-word verb meaning 'to increase' or 'to accumulate'.

The verb *run up* is one of thousands of multi-word verbs in the English language in which the particle is one of the following:

- always an adverb such as *away* in
  [1] The dog *ran away*.
- always a preposition such as *with* in
  [2] Will you *come with* me to the store?
- either an adverb or a preposition depending upon how it is used in a sentence. The following well-known examples illustrate this type of particle:
  [3] The mouse *ran up* the wall.
  [4] The soldiers *ran up* the flag (Cowie 1993).

The particle in [3] is a preposition while in [4] the particle is an adverb. *Up* in this sense is called an adprep because it can be used as either an adverb or a preposition.

The meaning of many, perhaps most, verb plus particle combinations can be deduced from the individual components as in the first three examples. The meaning of [4] is not so easily deduced; we must know that this sense of *run up* is synonymous with 'raise'. Verb plus particle combinations that can be replaced by a single verb are often idiomatic and the meaning of the combination cannot be deduced from the individual components. Consider these examples from Cowie:

[5] Someone in the office must have *put* that story *about*.
[6] The printer will *run off* three copies (Cowie 1993).

In these examples, *put about* can be replaced by the word 'circulate' while *run off* can be replaced by 'print'.

As we read the above statements, our minds go through a process of choosing the correct meaning for each sentence, often without being aware that such a process is taking place. Computer programs that interface with people, using natural language, must be able to perform the same disambiguation process. The MWCF database is designed to be a tool for researchers who are attempting to duplicate the above disambiguation process in their computer programs.

## The MWCF Database

The MWCF database is designed to serve dual functions. Figure 1 shows a dotted line to indicate the separation of the files as they relate to the multiple functions. The files above the dotted line, representing the Combined Case Frames, facilitate disambiguation research by associating all forms of a multi-word verb to case frames. The files below the dotted line are designed to provide a link between the case frames back to the data from the original dictionary databases. The following sections describe the database files.

### The FORM File

The verb is stored in its base form throughout the lexical database. The Form file allows the user to match any form of a verb to the base verb in the lexical database. The FORM File contains the following data entries:

- Mverb contains the verb plus particle combination for a given form of the verb.
- Verb contains the verb form. This field can be used to match the verb in a sentence to a verb in the lexical database.
- Part contains the particle. This field can be used to match particles in sentences to the lexical database.
- Hverb contains the head verb plus particle. This field is used to link to the Case Frame files in the lexical database.
- Verb Type contains the type of form. The codes are B1 for the base form, 3P for third person singular, Past for past tense, PrP for present participle and PaP for past participle.

### Combined Case Frames

Each verb sense can result in one or more case frames. In the CASES and CASELINES files, described later in this paper, at least one case frame was created for each sense of each verb plus particle combination that appears in either the *Collins English Dictionary* (Hanks 1979) or the *Collins COBUILD English Language Dictionary* (Sinclair 1987). To reduce the amount of redundant data, the cases were combined to create unique, combined occurrences of the case frames. Three files contain the combined case frames.

**The CCLine File.** Up to four verb arguments may be described for each verb, depending upon the usage of the verb. The CCLine file contains unique case line records containing data about each of the verb arguments. The fields in the CCLine file are:

- The CCLNum is a counter-generated field to uniquely identify the records.
- SynRole identifies the syntactic role of the argument. This field is used to indicate whether the argument is a subject, direct object, indirect object, or other sentence construct.
- Case identifies the case or thematic role of the argument. Allen's (1995) cases are used in this field.
- Occurrence indicates whether the argument is obligatory, optional, or elliptical.

**The CCase Frame File.** The CCase Frame file contains unique cases. The fields in the CCase Frame file are:

- VCNum is a counter-generated field to be used to uniquely identify the records.
- Entry contains the multi-word verb.

- The CCL fields, one for each verb argument (CCL1, CCL2, CCL3, CCL4) contain the CCLNum of the record in CCLine that contains the associated SynRole, Case, and Occurrence field.
- Pass is the passivity of the verb.

**The CCU Restr File.** The CCU Restr file contains the selectional-restriction fields for each combined case frame. Selectional-restriction fields contain data to indicate the restrictions placed on the verb arguments. For instance, selectional-restrictions for the object of the verb *run up* include clothing, debts, dress, flag and skirt. The CCU Restr file contains the VCNum, CLine, and Restr fields. CLine is the verb argument line number.

The three files can be linked together to obtain all the data required for a case frame. The combined case frames can also be used to create a Case Frame Report.

The report produced for the verb *run up* shows the following data for the first line:

- VERB: run up
- SYN-ROLE: subject
- CASE: agent
- OCCURRENCE: obligatory
- SELECT-RESTRICTION: human
- PASS: P

and the following data for the second line:

- VERB: run up
- SYN-ROLE: obj-direct
- CASE: theme
- OCCURRENCE: obligatory
- SELECT-RESTRICTION: clothing/debts/dress/flag/skirt
- PASS: P

## Case Frames by Word Sense

The files above the line in figure 1 were combined to eliminate redundant data. The files were created using Microsoft Access queries to produce unique rows using the CASES and CASELINES files. The CASES and CASELINES files contain one or more case frames for each verb sense in either dictionary. The CASES file contains information about the verb sense while the CASELINES file contains information about the verb arguments.

**The CASES File.** The fields in the CASES file include:

- CaseNum is a generated number to identify each case.
- BulkID is used to link the case frames back to the dictionary entry.
- CNum is used to uniquely identify cases in the event that there is more than one case frame per sense.
- Entry is the multi-word verb.
- MVType, used to indicate the multi-verb category, is used to indicate if the verb is a phrasal verb, prepositional verb, or a three-word phrasal-prepositional verb. The MVType field identifies the type of multi-word verb and can be used to assist in disambiguation during future research projects.
- PartType is used to indicate whether the particle is an adverb, preposition, or adprep.
- Passivity is used to indicate whether or not the verb can be used in the passive voice for this verb sense.
- Transitivity indicates whether the verb is transitive or intransitive for this verb sense.

**The CASELINES File.** The CASELINES file contains up to four lines for each CASES record. The CASELINES file is used to describe the verb arguments and contains the following fields:

- CaseNum used to associate the CASES and CASELINES files.
- CLine contains a line number, one line each for up to four arguments associated with the verb.
- SynRole identifies the syntactic role of the argument.
- Case identifies the case or thematic role of the argument.
- Occurrence indicates whether the argument is obligatory, optional, or elliptical.
- Each argument line contains four selectional-restriction fields that can be used to specify restrictions on the contents of the argument.

The case frames were automatically generated by Microsoft Access queries and then manually corrected.

## Auxiliary Files

Two auxiliary files were produced solely for the purpose of establishing links in the database.

**The CCase File.** The CCase file links the records in the CASES file to the records in the CCase Frame file. It contains the CaseNum field from CASES and the VCNum field from the CCase Frame file.

**The MVERB File.** The MWCF is designed to be incorporated into IITLEX which uses the machine-readable version of the *Collins English Dictionary* (Hanks 1979), also called the CED in this paper, as its primary source. While investigating the multi-word verbs prior to generating the databases, I determined that information included in the *Collins COBUILD English Language Dictionary* (Sinclair 1987), also called the COB in this paper, would be helpful in the creation of the case frames.

Using both dictionaries required an extensive manual match of all senses of multi-word verbs in the CED and the COB. The results of the match were stored in the MVERB file, the only file in the database that specifically links the verb senses from the two dictionaries. The fields in the MVERB file include:

- MVID uniquely identifies each record in the file.

- CED BulkID contains the identifier to link to the files containing data from the *Collins English Dictionary*.
- COB BulkNum contains the identifier to link to the files containing data from the *Collins COBUILD English Language Dictionary*.
- Entry contains the multi-word verb.
- CED Sense.
- COB Sense.
- Main Verb contains the verb portion of Entry.
- Particle contains the particle from Entry.

Each record in the MVERB File contains both the CED BulkID and the COB BulkNum. For a given sense of a verb, entries in both fields indicates a match. A zero in one of the fields indicates an unmatched verb.

The MVERB file contains 2089 multi-word verbs represented in 4370 verb senses.

## The CED Database

Strutz (1994) parsed the machine-readable version of the CED to create the CED BULKs. The BULKs for *run up* are shown below:

BEGIN_BULK:
BULK_ID: 1193933
VSEN: transitive
PARTICLE: adv
ENTRY: run up
WCNT: 2
POS: vb
SENSE: 1
TEXT: to amass or accumulate; incur
EXAM: to run up debts
END_BULK:

BEGIN_BULK:
BULK_ID: 1193934
VSEN: transitive
PARTICLE: adv
ENTRY: run up
WCNT: 2
POS: vb
SENSE: 2
TEXT: to make by sewing together quickly
EXAM: to run up a dress
END_BULK:

BEGIN_BULK:
BULK_ID: 1193935
VSEN: transitive
PARTICLE: adv
ENTRY: run up
WCNT: 2
POS: vb
SENSE: 3
TEXT: to hoist
EXAM: to run up a flag
END_BULK:

BEGIN_BULK:
BULK_ID: 1193936
ENTRY: run-up
WCNT: 1
HEAD: run up
POS: n
SENSE: 4
TEXT: an approach run by an athlete for a long jump, pole vault, etc.
END_BULK:

The BULKs were used to create a Microsoft Access database that contains all the information from the BULKs. The CED database contains four major files: CED1, CED2, CED3, and CEDTag.

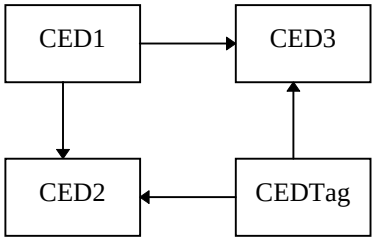


Figure 2. CED Database Files

The CED1 file contains fields that occur only once for each entry and that occur on most entries. The fields in the CED1 file are:

- BulkID is a unique identifier for each sense in the CED.
- Entry contains the multi-word verb.
- Homograph is the homograph number from the dictionary.
- Sense is the sense number for this entry in the dictionary.
- WCNT is the number of words in the Entry field.
- POS is the part-of-speech for this entry.
- Head is the head verb for this entry.

The CED2 and CED3 files contain fields that occur more than once per entry or that do not occur often among entries. The fields in the CED2 and CED3 files are:

- BulkID to link the record(s) to the corresponding record in the CED1 file.

- Tag Number to identify the type of data contained in each record.
- Tag Count to be used when multiple records exist for the same BulkID and Tag Number.
- Line Count to be used when more than one record is required to contain all the data for a BulkID, Tag Number, and Tag Count.
- Tag Text contains the contents of the dictionary entry for this tag..

The data in the CED2 and CED3 files are identified by a tag number that is associated with the CEDTag file. CED1, CED2, and CED3 are related through a field named BulkID which uniquely identifies each word sense. Both CED2 and CED3 have concatenated keys of BulkID, Tag Number, Tag Count, and Line Count.

The CEDTag File is used to keep track of the types of information stored in CED1, CED2, and CED3. The CEDTag file contains:

- Tag Number identifies the record and is used to link the record to the CED2 or CED3 file.
- Database indicates whether the data is in CED1, CED2, or CED3.
- Tag Text contains the data from the dictionary entry.

## Case Frame Generation

The MWCF database was created in the reverse order of the above description of the database files. The following paragraphs briefly describe the process. For additional details, see Kowalke (1998).

The first step was to generate the CED database from the BULK flat files using a C formatting program. The resulting files were then imported into Microsoft Access to create the CED database files. The COB database was created using a similar method.

A manual matching of the CED and COB multi-word verbs at the sense level was conducted. The results of the match were used to create the MVERB file. This step was by far the most time-consuming task. However, the resulting set of multi-word verbs with the associated data produced a more comprehensive database.

The CASES and CASELINES files were created next in two phases. Using Microsoft Access queries, the files were automatically generated from the data in the two databases and the MVERB file. The most difficult task was the generation of the verb argument data in the CASELINES file. The following paragraphs briefly summarize the creation of the CASELINES file.

Since a subject is generally required for all sentences, all entries for the subject argument on line one were given the syntactic role of subject, a default case of agent, and a default occurrence of obligatory.

The object line defaults were more difficult to generate. For example, if the COB grammar note is V + ADV, no additional lines were required. However, grammar notes that include V + ADV + O or V + O + ADV require a second line. The defaults for the direct object argument were obj-direct for syntactic role, theme for case, and obligatory for occurrence. If the grammar note includes V + PREP of V + ADV/PREP, the syntactic role was assigned as obj-prep, the case field as theme, and the occurrence as obligatory. The role of obj-prep was temporary and was adjusted to the preposition plus "-np" during the manual correction phase.

An adjunct is a noun group, a prepositional group, or an adverbial group that expresses time, place, manner, or condition. Determining which line to automatically generate depended upon whether or not an object line had been generated. If line two had already been generated for the direct object, the adjunct was placed in line three. If no object had been generated, the adjunct was placed in line two.

Many verb senses required the creation of more than one case frame. For example an ergative verb can be both transitive and intransitive in the same meaning and requires two case frames. Case one was generated for the transitive usage and contains a subject and an obj-direct line while case two was generated for the intransitive usage and requires only a subject line. For case one the generated values for line one were subject for syntactic role, agent for case, obligatory for occurrence. The generated values for line two were obj-direct for syntactic role, theme for case, and obligatory for occurrence. For case two, the generated values for line one were subject for syntactic role, theme for case, and obligatory for occurrence.

The automatically-generated case frames were then manually corrected. After manual correction, the CASES and CASELINES files were processed by Microsoft Access queries to produce unique case frames which can be used for word sense disambiguation. Finally, the CCase file was created to link the CASES and CCase Frame file.

## Conclusion

The MWCF database is designed as a tool to be used by future word-disambiguation projects to determine if a verb plus particle combination is a multi-word verb. This paper has described the database and briefly summarized the process used to generate the database. Throughout the entire generation process, special care was taken to create and preserve the links between the files to allow the user to trace the use of a multi-word verb through the case frames back to the dictionary databases. This flexibility will make all of the data related to a multi-word available to the researcher using the database.

As a preliminary test of the database, the combined case frame files, those above the line in figure 1, were used to determine if free combinations of a verb plus particle could

be syntactically eliminated. As an example, of 47 sentences from the *Wall Street Journal* that contain both the word run and the word up, 24 were eliminated as free combinations. Of the remaining 23 sentences, all but two were phrasal verbs.

Copies of the "above the line" case frame files will be available on request after the conference. Hard copy versions of the case frame report can also be obtained.

## References


Allen, J. 1995. *Natural Language Understanding (Second Edition)*, Menlo Park, CA: Benjamin Cummings Publishing Company, Inc.

Cowie, A. 1986. Strategies for Dealing with Idioms, Collocations, and Routine Formulae in Dictionaries. Paper presented at *Grosseto Workshop on the Lexicon*.

Evens, M., Dardaine, J., Huang, Y., M. Li, S., Markowitz, J., Rinaldo, F., Rinaldo, M., Strutz, R. 1991. For the Lexicon That Has Everything. *Proceedings of the Siglex Workshop*, Berkeley, CA, June, 1991. 179-187. An extended version appeared in J. Pustejovsky and S. Bergler, Eds. *Lexical Semantics and Knowledge Representation.* Berlin, Germany: Springer-Verlag. 219-233.

Hanks, P., Ed. 1979. *Collins English Dictionary*. Birmingham: Collins Publishing.

Kowalke, P. M. 1998. *Incorporation of Entries for Phrasal Verbs Into a Lexical Database*, Ph.D. diss., Computer Science Department, Illinois Institute of Technology.

Sinclair, J., Ed. 1987. *Collins COBUILD English Language Dictionary*. London, UK: HarperCollins Publishers.

Strutz, R. E. 1994. *Construction of a General Purpose Lexical Database and Access Tool*. Ph.D. diss., Computer Science Department, Illinois Institute of Technology.