Hence there is a real need for a utility to validate data *names* – effectively a CIF spelling checker.

5.3.4.1. *CYCLOPS*

The program *CYCLOPS* (Hall, 1993; Bernstein & Hall, 1998) was written specifically to address the problem of validating CIF data names. Its use extends beyond simply identifying data names in a CIF data file and checking that they are defined in a dictionary. Any ASCII file may be input, allowing for the checking of CIF data names in any text documents or program source.

The program was originally written in Fortran as an aid to ensuring that the original core CIF dictionary was free from data-name errors; subsequently it was extended to be able to read multiple dictionaries in DDL1 and DDL2 formats, and to resolve data-name aliases across multiple dictionaries. The extended version was written with the library routines of the *CIFtbx* toolkit (Hall & Bernstein, 1996) described in Chapter 5.4 and is distributed as an example application with *CIFtbx*. The description below refers to this extended version, also known as *CYCLOPS2*.

5.3.4.1.1. *Operation*

The program determines the dictionary (or list of dictionaries) against which to validate the input text file (see below for the method of passing such information to the program). It opens each dictionary in turn and stores all data names defined in the dictionaries. Where the same name is defined in multiple dictionaries, the behaviour is determined by a command-line switch.

The text file is then input and parsed for candidate data names. Because the program is designed to check potential data names embedded in ordinary text files, it is not sufficient to apply the CIF parsing rule of a white-space-delimited character string beginning with an underscore character. Instead, character strings are sought that begin with an underscore optionally preceded by white space or one of the characters `,.([{</\|'":*` and followed by white space, one of the characters `,.)]}>/\|'"-=?!;:` or by the end of a line.

For each candidate data name found in this way, matching data names in the stored list are identified in one of three ways:

(i) If the data name is not preceded by the asterisk character `*` and it does not end with the underscore character _, then search for an identical match.

(ii) If the data name ends with the underscore character _, then search for a match in the dictionary where the leading characters in the dictionary name are the same as all the characters in the data name found in the text. For example, the text `_atom_site.label_` would match the mmCIF dictionary entry **`_atom_site.label_alt_id`**.

(iii) If the data name is preceded by the asterisk character `*`, then search for a match in the dictionary where the trailing characters in the dictionary name are the same as all the characters in the data name found in the text. The first match found in the dictionary is accepted. For example, the text `*_alt_id` would match **`_atom_site.label_alt_id`**, or, if that name had not been in the dictionary, **`_struct_conn.ptnr1_label_alt_id`**. If one of the searches succeeds, add the line number of the data name to a list attached to the dictionary name. Up to 19 line numbers are retained for each dictionary name (the first ten matches and the last nine).

If no match is found, the unmatched data name is added to the list of unmatched names, along with the appropriate line number. If a data name has been misspelled it will be caught at this step.

When the text file has been processed, a validation report file is output containing the alphabetically sorted list of unmatched names and line numbers, followed by the sorted list of names from all dictionaries that are used within the text. If requested, this is followed by the sorted list of names from all dictionaries that are not used within the text in the file. If a data name has an alias defined in the dictionaries, a warning about the existence of the alias is given. If more than one dictionary has been used, the source dictionary is identified for each data name. An example of the output from *CYCLOPS* is shown in Fig. 5.3.4.1.

5.3.4.1.2. *Invocation of the program*

CYCLOPS is generally invoked from a command line that specifies the input and output file names and the dictionary files against which to validate the input. However, because the program is portable across a wide range of operating systems, there is substantial flexibility in the way in which it may be invoked. Under a Unix-like operating system, the program may typically be called with a command such as

`cyclops -i infile -o outfile -d dictfile`

where *infile* is the name of the input file for validation, *outfile* is the file to which the detailed output of the program is written and *dictfile* is a dictionary file.

A more complete set of options available in a Unix-like operating environment is

`cyclops [-i infile] [-o outfile] [-d dictfile] [-p priority]`
`[-f cmndfile] [-c catck] [-v verbose] [-s short]`

where the options are as follows:

-i specifies the name of the input file, *infile*.

-o specifies the name of the output file, *outfile*.

-d specifies the name of the dictionary file, *dictfile*. For compatibility with the original version of the software, the dictionary file may be *either* a CIF dictionary or a list of file names. That is, it may contain dictionary definitions in DDL format or (if the file begins with the characters `#DICT`) it may contain a list of dictionary file names to be entered. As implied by this last statement, multiple dictionaries may be specified to the program.

-p specifies the priority that should be assigned if multiple definitions for the same data name are encountered when multiple dictionaries are accessed. The permitted values are: *first* (the default), in which the first of duplicate definitions to be loaded takes priority; *final*, in which the last takes priority; and *nodup*, in which an instance of a duplicate definition should be treated as a fatal error.

-f specifies the name of a command file *cmndfile* that contains additional directives to the program.

-c is a flag indicating whether an error message should be raised if a data name has been assigned a category different from the leading portion of the data name itself. The Boolean variable *catck* may take the values 't', '1' or 'y' for *true*, 'f', '0' or 'n' for *false*.

-v is a flag indicating whether a verbose listing of unreferenced data names should be generated. The Boolean variable *verbose* may take the same values for *true* or *false* as above.

-s is a flag indicating whether the output should be short (*i.e.* restricted to items not in dictionaries). The Boolean variable *short* takes the same values as above.

For the flags expecting Boolean values, the default is 'f' (*false*).

If no input or output file names are specified, the program will read from the standard input channel or write to standard output,