User Manual

UCC v.2013.04B

Version History

Date	Author	Version	Changes
01/23/2007	Marilyn Sperka	0.1	Initial Version
01/30/2007	Vu Nguyen	0.3	Reviewed and updated installation procedures and terminologies to ensure they are consistent with the readme file.
02/03/2009	Vu Nguyen	0.4	Added an instruction to include –DUNIX in the compile line on Unix/Linux.
10/26/2009	Vu Nguyen	1.0	Updated for Release 2009.10
01/12/2010	Marilyn Sperka	1.1	Reformatted and minor edits
06/10/2010	Marilyn Sperka Vu Nguyen	2.0	Updated for UCC v.2010.06
03/10/2011	Marilyn Sperka	2.3	Updated for UCC v 2011.03
05/02/2011	Marilyn Sperka	2.4	Updated for UCC v 2011.05
10/28/2011	Marilyn Sperka	2011.10B	Updated for UCC v2011.10B (B=Beta release)
04/30/2012	Marilyn Sperka Ryan Pfeiffer	2011.10	Updated for UCC v2010.10 release
03/28/2013	Ryan Pfeiffer	2013.04B	Updated for UCC v2013.04B

Table of Contents

1 Introduction

This document provides information for using the UCC tool version 2013.04B.

1.1 Product Overview

Most software cost estimation models including the COCOMO model require some sizing of software code as an input. Ensuring consistency across independent organizations in the rules used to count software cost code is often difficult to achieve. To that end, the USC Center for Systems and Software Engineering (CSSE) has developed and released a code counting toolset called CodeCount to support sizing software code for historical data collection, cost estimation, and reporting purposes. This toolset is a collection of tools designed to automate the collection of source code sizing information. It implements the popular code counting standards published by SEI [1] and adapted by the COCOMO model [2]. Logical and physical source lines of code (SLOC) are among the metrics generated by the toolset.

Unified CodeCount (UCC) is a unified and enhanced version of the CodeCount toolset. It is a code counting and differencing tool that unifies the source counting capabilities of the previous CodeCount tools and source differencing capabilities of the Difftool (which is now replaced by UCC). It allows the user to count, compare, and collect logical differentials between two versions of the source code of a software product. The differencing capabilities allow users to count the number of added/new, deleted, modified, and unmodified logical SLOC of the current version in comparison with the previous version. With the counting capabilities, users can generate the physical and logical SLOC counts, and other sizing information such as complexity, comment and keyword counts of the target program.

The UCC tool is provided in C++ source code, and may be used as is, or modified and further distributed subject to certain limitations. The user is responsible for compiling and using the executable version.

2 System Requirements

Note: with large files or baselines, long run time and hanging may be experienced. Refer to the section Large_Jobs for performance issues.

2.1 Hardware

- RAM: 512 MB. Recommended: 1024 MB.
- HDD: 100 MB available. Recommended: 200 MB available.

2.2 Software Operating Systems

- Linux
- Unix
- Mac OS X
- Windows XP/7
- Solaris

2.3 Compilers Supported

- MS Visual Studio 2008, 2010, 2012
- MinGW
- g++
- Eclipse C/C++
- Any ANSII-Standard C++ compiler RAM: minimum 512 MB. Recommended: 1024 MB

3 Installation

There is no setup package provided for installing the tool.

3.1 Software Download

The user can download the source files from http://csse.usc.edu/ucc.

3.2 Compilation

The tool can be compiled using an ANSII-Standard C++ compiler. The compiler must support common C++ libraries including IO and STL. Below are typical steps for compiling the tool using Visual Studio, MinGW, and the g++ compilers. The procedure for compiling the tool using Eclipse C/C++ would be the same as typical C++ programs.

3.2.1 Visual Studio

UCC is a command-line application and it must be compiled under the application console mode. On PC based machines, the user can use Visual Studio 2008, 2010, or 2012 to compile the source code by following the procedure:

1) Create an empty project of Project Type Visual C++ and Template Win32 Console Application. Type in the Project name and Select OK. In the Win32 Application Wizard, select Applications Settings and then select “Empty Project” check box.

2) Select Project/Add Existing Item. Locate and select all UCC source code files. Click on “Add” to add the selected files. This would add the existing code to the created project.

3) Open the Properties window and go to the Configuration mode page. The user should select "Release" mode for compiling. To choose the mode, Click on Build/Configuration Manager button. Select "Release" from the list "Active Solution configuration". Go to the C/C++ section and click "Precompiled Headers". Make sure the "Create/Use Precompiled Header" selection is "Not Using Precompiled Headers". (*)

4) Select Build Solution or use the shortcut Ctrl+Shift+B to compile.

Upon compilation, an executable file will be created in the *Release* folder.

(*) Please note that Visual Studio by default uses the precompiled header option. The user will get error messages if this option is not turned off.

3.2.2 MinGW

The following command will compile source files stored in the folder *src*.

g++ ./src/*.cpp -o UCC -DMINGW

Note: the command line option -DMINGW must be provided when using the MinGW compiler.

3.2.3 g++

The following command will compile source files stored in the folder *src*.

g++ ./src/*.cpp -o UCC -DUNIX

Note: the command line option -DUNIX must be provided on UNIX-based systems, including Cygwin and Mac OSX.

4 Execution

4.1 Command Line Specification Summary

This section describes the command line format used to invoke UCC, along with a summary of the various switches and their usage. Section 7 will describe each switch in detail along with information on how to tailor the switch usage for various execution requirements as well as performance implications.

UCC [-v] [-d [-i1 fileListA.txt] [-i2 fileListB.txt] [-t #]] [-tdup #] [-trunc #] [-cf] [-dir <dirA> [dirB] <filespecs>] [-outdir outDir] [-extfile extFile] [-unified] [-ascii] [-legacy] [-nodup] [-nocomplex] [-nolinks]

-v Displays the version number of the UCC being executed.

-d If specified, UCC will run the differencing function. Otherwise, the counting function is executed.

-i1 fileListA.txt — Input list filename containing filenames and/or directories in *Baseline* A to be counted if -d is not present or compared if -d is present. The file is in plain text format, one filename or directory name per line. If a directory name is specified, the directory is searched recursively for all countable files.

-i2 fileListB.txt — Input list filename containing filenames in *Baseline* B. If the -d switch is present, the differencing function will be invoked and fileListA.txt and fileListB.txt will be compared. Normally, *Baseline* B is the newer or the current version of the program, as compared to *Baseline* A. The file format is same as *Baseline* A.

-t # — Specify the modification threshold, the percentage of common characters between two lines of code to be compared over the length of the longest line. If two lines have the percentage of common characters equal or higher than the specified threshold, they are matched and counted as modified. Otherwise, they are counted as one SLOC deleted and one SLOC added. The valid values range from 0 to 100 and default to 60 (same as *-t 60*).

-tdup # — Specify the threshold percentage for duplicated files of the same name. This specifies the maximum percent difference between two files of the same name in a baseline to be considered duplicates. By default, this threshold is zero, so the files must be identical by logical SLOC – spacing and comments are not considered.

-trunc # — Truncate threshold, specifying the maximum number of characters allowed in a logical SLOC. Additional characters will be truncated. The default value is 10,000, and zero is for no truncation. Performance can be significantly degraded if truncation is too high.

-cf — Support handling ClearCase filenames. The ClearCase application appends version information to the filename, starting from '@@'. This option requires the UCC tool to handle the original filename instead of the ClearCase-modified filename.

-dir — Specify directories containing files to be counted and/or compared. If this argument is provided, the input list files (see above) are ignored. If the –d is not present, UCC looks for dirA and filespecs. If the -d is present, UCC looks for dirA, dirB and filespecs. The directories are searched recursively for files that match the filespecs. See section 3.2.1.1 for examples.

dirA — Name of the directory. If the option -d is provided, dirA is the directory of *Baseline* A. Otherwise, it specifies the directory to be counted with the counting function.

dirB	Name of the directory of *Baseline* B, used only if the option -d is given.
filespecs	Specifications of file extensions to be counted/compared; wildcard chars ? * are allowed. Use a space between *filespecs*; for example, *.cpp *.c
-outdir <dirname>	Allows user to specify a directory name to be prepended to the output reports. The reports will be generated in the specified directory. This allows the user to set up a batch job with multiple UCC runs and not have the output reports overwritten.
-extfile <extfile>	Allows user to specify the name of a file that contains languages and extensions to map to that language counter. This allows the user to include non-default extensions, or remove default extensions. See section 6.1 for examples.
-unified	Directs the UCC to print counting results to a single unified language file name TOTAL_outfile.csv.
-ascii	Generate reports in text format. Default is .csv.
-legacy	Generate reports in legacy format. Use this option to retain the output files formats supported by Difftool, CodeCount Tools Release 2007.07 and earlier versions. By default, the new output files formats with new fields are applied.
-nocomplex	Do not process language keywords or report complexity metrics. Using this switch will reduce the processing time.
-nodup	Do not search for duplicate files. Any duplicates will be reported as unique files. This decreases processing time.
-nolinks	Skip Unix symbolic links to prevent multiple counting of same files as links.

Note: If no argument is given the tool reads and counts all files listed in the text file *filelist.txt or filelist.dat*. Details on this file are described in Section 4.2.1.2.

4.2 Counting and Differencing Details and Examples

4.2.1 Counting Source Files

The counting function is executed using the command line with no *-d* switch. There are two alternative ways to specify the source files of the target program: using the *-dir* switch and using the file list (*fileList.txt*).

4.2.1.1 Using the –*dir* command line switch

This command requires the tool to count all source files contained in the folder *project1*:

UCC -dir project1

This command requires the tool to count all C/C++ source files contained in the folder *project1*:

UCC -dir project1 *.cpp *.h *.c *.hpp *.cc

This command requires the tool to difference all source files contained in the folder *project1* with those contained in *project2*:

UCC -d -dir project1 project2

This command requires the tool to difference all C/C++ source files contained in the folder *project1* with those contained in *project2*:

UCC -d -dir project1 project2 *.cpp *.h *.c *.hpp *.cc

Under Unix/Linux when using the –dir option, any wildcards must be enclosed within quotes. Otherwise, the wildcards will be expanded on the command line and erroneous results will be produced. For example: ucc -d -dir baseA baseB *.cpp should be written as ucc -d -dir baseA baseB "*.cpp".

4.2.1.2 Using *fileList.txt*

UCC

This command requires the tool to find the file *fileList.txt* in the working directory and count all source files listed in it.

The file *fileList.txt* contains a list of source files to be counted, one filename per line. You can create the file *fileList.txt* using one of the following commands

- Unix:

  ls -1 <filespecs> > fileList.txt

  or, use the following to obtain full directory/pathname specification

  find *[Directory]* -name '<filespecs>' > fileList.txt

  to append this file with additional filenames use:

  find *[Directory]* -name '<filespecs>' >> fileList.txt

- MS-DOS:

  dir/B > fileList.txt *[Directory]*\<*filespecs*>
  or, use the following to obtain full directory/pathname specification

  along with files in all subdirectories:

  dir/B/S > fileList.txt *[Directory]*\<*filespecs*>

  to append this file with additional filenames use:

  dir/B/S > fileList.txt *[Directory]*\<*filespecs*>

  Where,

  - *Directory* is the directory name relative to the current working directory. *Directory* is optional, and it is not given, the working directory is implied.
  - *filespecs* is the file specifications, and wildcard chars ? * are allowed. Use a space between two filespecs, for examples, *.cpp *.c

4.2.1.3 Specify a filename containing the input list of files

UCC -i1 <fileListA>

This command requires the tool to find the file <*fileListA.txt*> in the working directory and read it to obtain the input source files listed in it. Since the format of these files is the same as *fileList.txt*, you can use the commands described in 4.2.1.2 to create them.

An advantage to this method is that the input file list can be given a descriptive name, such as JustCppFiles.txt, and multiple input file lists can be stored in the same directory.

4.2.2 Differencing Baselines

In this function, source files in the baselines will be matched and compared to determine the counts for SLOC added, deleted, modified, or unmodified.

To run this function, UCC must be called with the *-d* switch.

4.2.2.1 Using List Files

Compare source files of Baseline A and Baseline B contained in files *fileListA.txt* and *fileListB.txt.*

UCC -d

By default, the files *fileListA.txt* and *fileListB.txt* contains source filenames in *Baseline* A and *Baseline* B, respectively. You can specify different filenames by using the *-i1* and *-i2* command line switches.

UCC -d -i1 fileA.txt -i2 fileB.txt

(Since the format of these files is the same as *fileList.txt*, you can use the commands described in 3.2.1.2 to create them).

4.2.2.2 Using the *-dir* command line switch

UCC -d -dir <dirname1> <dirname2> filespec1 filespec2 … filespecn

For example:

UCC -d -dir code1.0 code1.2 *.cpp *.c *.hpp *.h *.cc

requires the tool to match and compare all source files with extensions of *.cpp, *.c, *.hpp, *.h, or .cc contained in the folder *code1.0* and *code1.2*.

5 Output Files

A variety of output files are produced in order to meet the needs of different types of users. The reports are by default produced in .csv format which can be opened using Excel where further analysis can be done by the user. The reports can be produced in text format by including "-ascii" on the command line, and the files will have the extension .txt.

<LANG> is the name of the language of the source files, e.g., C_CPP for C/C++ files and Java for Java files.

File Name	Function	Description
error_log_<mmddyyyy>_<hhmmss>	Both	Log file listing errors that occur at the date specified by month, date, year <mmddyyyy> and time specified by hours, minutes, seconds <hhmmss>
<LANG>_outfile.csv	Counting	Counting results for source files of <LANG>
outfile_summary.csv	Counting	Summary counting results for all languages and source files counted
outfile_cplx.csv	Counting	Complexity results
outfile_cyclomatic_cplx.csv	Counting	Cyclomatic complexity results
Duplicates-outfile_cplx.csv	Counting	Complexity results for duplicate files when counting
Duplicates-outfile_cyclomatic_cplx.csv	Counting	Cyclomatic complexity results for duplicate files when counting
Duplicates-<LANG>_outfile.csv	Counting	Counting results for duplicate files of <LANG>
DuplicatePairs.csv	Counting	Lists original and duplicate file pairs
outfile_diff_results.csv	Differencing	Main differencing results
MatchedPairs.csv	Differencing	Shows how files in Baseline A and Baseline B were paired for differencing
Baseline-<A|B>-<LANG>_outfile.csv	Differencing	Counting results for source files of <LANG> for *Baseline* A and *Baseline* B
Baseline-<A|B>-outfile_cplx.csv	Differencing	Complexity results for *Baseline* A and *Baseline* B
Baseline-<A|B>-outfile_cyclomatic_cplx.csv	Differencing	Cyclomatic complexity results for *Baseline* A and *Baseline* B
Baseline-<A|B>-outfile_summary.csv	Differencing	Summary counting results for all languages and source files in *Baseline* A and *Baseline* B
Duplicates-<A|B>-<LANG>_outfile.csv	Differencing	Counting results for duplicate files of <LANG> in *Baseline* A and *Baseline* B
Duplicates-<A|B>-outfile_cplx.csv	Differencing	Complexity results for duplicate files in *Baseline* A and *Baseline* B
Duplicates-<A|B>-outfile_cyclomatic_cplx.csv	Differencing	Cyclomatic complexity results for duplicate files in *Baseline* A and *Baseline* B
Duplicates-<A|B>-DuplicatePairs.csv	Differencing	Lists original and duplicate file pairs in *Baseline* A and *Baseline* B
Duplicates-A-outfile_summary.csv		Summary counting results for duplicate files for all languages in *Baseline* A and *Baseline* B

outfile_uncounted_files.csv	Both	Lists files that were unable to be counted along with the reason (if known)

6 Counting Standards

The UCC counts physical and logical SLOC and other metrics according to published counting standards which are developed at CSSE so that the logic behind the metrics being produced is clear to all participants. The counting standard documents are separate documents and are included in the UCC release. The counting standards are derived from the latest available ANSII standard language specification for each language counted by the UCC. Users should note that non-ANSII standard compilers may have commands which are outside of the ANSII standard specification. The results from using UCC on non-ANSII standard code cannot be guaranteed.

The counting standards documents for each language provide detailed information of what is counted, and how items are counted, so that all users can understand the operations and outputs of the UCC. Definitions are included of what is considered to be a blank line, comment line, and executable line of code for each language. The document describes in detail the physical and logical SLOC counting rules. Physical SLOC are counted at one per line. Logical SLOC counting rules are grouped by structure and the order of precedence is defined.

The items being measured, in order of precedence (numbered), are:

- 1. Executable lines,
- Non-executable lines
    - 2. Declaration (Data) lines
    - 3. Compiler directives
    - Comments
        - 4. On their own lines
        - 5. Embedded
        - 6. Banners
        - 7. Empty comments
        - 8. Blank lines

A table of logical SLOC counting rules is provided, and further specifies the order of precedence for the various types of executables lines. The rules define precisely when a count occurs, and a comments section gives further explanation.

The counting standards define for each language what keywords are counted and tallied in the output report. The keywords include compiler directives, data keywords, and executable keywords. Compiler directives are statements that tell the compiler how to compile a program but not what to compile. Data keywords define data declarations, which describe storage elements and the format that will be used to interpret data contained in them. Executable keywords are execution control statements. The specified compiler directives, data keywords, and executable keywords are counted and included in output reports. The data keywords are specific to each language.

7 Terminology Explanation

7.1 File Extensions

The tool determines the language in a source file using its file extension. This version supports the following languages and file extensions:

Languages	File Extensions
Ada	.ada, .a, .adb, .ads
ASP, ASP.NET	.asp, .aspx
Bash	.sh, .ksh
C Shell Script	.csh, .tcsh
C#	.cs
C/C++	.cpp, .c, .h, .hpp, .cc, .hh
ColdFusion	*.cfm, .cfml, .cfc
ColdFusion Script	.cfs
CSS	.css
Data	Use file mapping with Datafile=<ext>
Fortran	.f, .for, .f77, .f90, .f95, .f03, .hpf
HTML	.htm, .html, .shtml, .stm, .sht, .oth, .xhtml
Java	.java
JavaScript	.js
JSP	.jsp
Makefiles	.make, .makefile, (files named Makefile)
MATLAB	.m
NeXtMidas	.mm
Pascal	.pas, .p, .pp, .pa3, .pa4, .pa5
Perl	.pl, .pm
PhP	.php
Python	.py
Ruby	.rb
SQL	.sql
VB	.vb, .frm, .mod, .cls, .bas
VBScript	.vbs
Verilog	.v
VHDL	.vhd, .vhdl
X-Midas	.txt
XML	.xml

It may be desirable to associate an additional extension to a language counter, or to disassociate a particular extension from a language counter. This can be done using the -extfile <filename> option on the command line. For more information, see section 8.11.

7.2 Basic Assumption and Definitions

7.2.1 Data Files

Data files shall contain only blank lines and data lines. Data lines are counted using the physical SLOC definition.

7.2.2 Source Files

Source code files may contain blank lines, comment lines (whole or embedded), compiler directives, data lines, or executable lines. Source code files have to be compiled successfully to ensure the integrity of the inclusive syntax.

7.2.3 SLOC Definitions and Counting Rules

Please refer to the counting standard documents.

7.2.4 TAB

A Tab character is treated as a blank character upon input.

7.2.5 Blank Line

A blank line is defined as any physical line of the source file that contains only blank, Tab, or form feed characters prior to the occurrence of a carriage return (EOLN).

7.2.6 Total Sizing

The total sizing of analyzed source code files in terms of the SLOC count contains the highest degree of confidence. However, the sizing information pertaining to the sub classifications (compiler directives, data lines, executable lines) has a somewhat lower level of confidence associated with them.

Misclassifications of the sub classifications of SLOC may occur due to:

1) user modifications to the UCC tool,
2) syntax and semantic enhancements to the parsed programming language,
3) exotic usage of the parsed programming language, and
4) integrity of the host platform execution environment.

Additionally, in some programming languages a single SLOC may contain attributes of both a data declaration and an executable instruction simultaneously. These occurrences represent events beyond the control of the UCC tool designer and may cause the inclusive parsing capabilities of the tool to misclassify a particular SLOC. For these reasons, the counts of sub-classifications should be regarded as an approximation and not as a precise count. In only the physical SLOC definition does the sum of the sub-classification counts equal the total physical SLOC count.

7.2.7 Keyword Count

The search for any programming language specific keywords over a physical line of code for purposes of incrementing the tally of occurrences shall include the detection of multiple keywords of the same type, e.g., two occurrences of the keyword READ on the same physical line.

The search for any programming language specific keywords over a physical line of code for purposes of incrementing the tally of occurrences shall include the detection on multiple keywords of different types, e.g., occurrences of keywords READ and WRITE on the same physical line.

Keywords found within comments (whole or embedded) or string literals shall not be included in the tally count.

8 Switch Usage Detail

This section will describe each switch in detail along with information on how to tailor the switch usage for various execution requirements. When applicable, performance implications will be described as well.

8.1 -v

Displays the version number of the UCC being executed.

Initially, the code counting tools were individual counters, and a separate program did differencing. These programs are still available on the public website at **http://csse.usc.edu/ucc** and are named the CodeCount Tools – Release 2007.07. These tools were combined into a monolith program titled the Unified CodeCount (UCC). The public website offers multiple versions of the UCC. The UCC Title indicates the year and month of the release. For example, Release 2009.10 was released in October of 2009. Users may choose to use previous releases for reasons which may include there being a bug in a more current release, or a bug was fixed in the current release and the user wants consistency with the way source was previously counted. For releases dated 2011.03 and later, including -v on the command line will cause UCC to print the version number to the standard output device. The message format is "UCC version 2011.03", meaning UCC release 2011.03, or the release of March, 2011.

8.2 -d

If specified, UCC will run the differencing function. Otherwise, the counting function is executed.

The UCC can be used to count SLOC within files, but it can also be used to difference two baselines of files. When just counting and not differencing is desired, use the UCC command without the -d switch. Counting is the default. When differencing, counting is performed and reported, and additional reports are produced which compare the files in two baselines and determine, for each file, how many lines of code were added, deleted, modified, or not modified. For counting and differencing, use the UCC -d command.

Differencing is a powerful capability that allows code changes between two baselines (or versions) of code to be monitored. Both physical and logical SLOC are included to provide insight into the extent of work completed between baselines.

The differencing process matches files between the two baselines using an algorithm that ensures the best possible match is found for all files. The matched file pairs are compared to each other line by line to determine how many SLOC have been added, deleted, modified, or are unmodified. All SLOC from any files in Baseline A that are not matched to a file in Baseline B are considered deleted, and all SLOC from any files in Baseline B that are not matched to a file in Baseline A are considered added.

The user is given the ability to tailor how the UCC determines if a SLOC has been modified using the -t # switch. For more information, see section 7.5.

The following relationships hold:

Baseline A SLOC = Deleted SLOC + Modified SLOC + Unmodified SLOC

Baseline B SLOC = New SLOC + Modified SLOC + Unmodified SLOC

Differencing will add a significant amount of time to the processing. It also will require more memory. If large numbers of files are in each baseline, there is a possibility that memory will become completely used and the process will hang. This problem may be addressed in a variety of ways.

The user may be able to run the process on a computer that has more memory than the computer that hung, or they may be able to add memory to the computer that hung.

The user may divide the input files into a number of smaller sets. If the user orders the smaller sets by language types, the duplicate file function will still work appropriately. If the files of a language type must be placed in separate sets, the duplicate file function may not find all duplicate pairs. If multiple sets are used, the user may need to aggregate the output files, as this will not be done automatically. The -outdir switch may be used to direct the outputs of sequential UCC executions to different directories, enabling the user to more easily specify multiple UCC runs. See section 8.10 for more information.

The user may choose to disable the duplicate file process by adding the -nodup switch to the UCC command line. The search for duplicate files will not be performed, resulting in a performance speedup. This may also result in a reduction of the amount of memory needed, so larger file sets may be possible.

Algorithms which may contribute to the ability of the UCC to avoid memory problems are being explored and may be included in future releases.

8.3 -i1 fileListA.txt

Input list filename containing filenames in *Baseline* A.

The fileListA.txt file is required to be a plain text format file containing a list of filenames to be processed, one per line. The files may be specified as full or relative directory/pathname specification. The UCC will open the file specified by the fileListA.txt argument, read each line and attempt to open each file specified. If a file cannot be opened, an error message is generated, and a tally is kept and reported in the output report.

When counting, as specified by the lack of -d switch on the command line, the UCC will expect just the -i1 switch and not the -i2 switch. When differencing, as specified by the inclusion of the -d switch on the command line, the UCC will expect to find the -i2 switch. If these conditions are not met, an error is generated.

8.4 -i2 fileListB.txt

Input list filename containing filenames in *Baseline* B.

If the -d switch is present, the differencing function will be invoked and fileListA.txt and fileListB.txt will be compared. Normally, *Baseline* B is the newer or the current version of the program, as compared to *Baseline* A. The file format is same as *Baseline* A.

8.5 -t

Specifies the modification threshold.

The user is given the ability to tailor how the UCC determines if a SLOC has been modified using the –t # switch. The # is the modification threshold and can be any integer between 0 and 100, with the default being 60 (same as -t 60). The modification threshold specifies a percentage; i.e. -t 60 indicates a modification threshold of 60%. If two SLOC have the percentage of common characters equal or higher than the specified threshold, as compared over the length of the longest line, they are matched and counted as modified. Otherwise, they are counted as one SLOC deleted and one SLOC added. In this example, using -t 60, if 60% of the characters of the longest line match with the compared line, the lines are considered modified in Baseline B. If less than 60% of the characters in the longest line match with the compared line, Baseline A counts one SLOC deleted, and Baseline B counts one SLOC added. If the compared lines are exactly the same, the lines are counted as unmodified.

8.6 -tdup

Specifies the threshold percentage for identical SLOC when comparing two files within a baseline for one file to be considered a duplicate of the other.

This switch specifies the maximum percent difference allowed between the SLOC in two files of the same name within a baseline to be considered duplicates. Blank lines and comments are not

considered. By default, this threshold is zero, so the files must be identical by logical SLOC in order to be considered duplicates. Valid values are integers between 1 and 100. The integer corresponds to the percentage of SLOC which may be different when two files are compared in order to be considered duplicates. For example, -tdup 20 would mean that a file is a duplicate of another file if 20% or less of the SLOC are different.

Duplicate file processing is computationally expensive. A switch -nodup is available to inhibit the duplicate processing in order to reduce execution time.

Files must be in the same baseline, but not necessarily in the same directory, in order to be duplicates. The files do not need to have the same name; however, if the file names are different, they have to be exactly the same, including comments and blank lines, to be identified as duplicates. Files with the same name, but not in the same subdirectory, are considered duplicates if the code is identical, even if there are differences in the comments and/or blank lines.

Duplicate files are counted and reported separately. One purpose for this command is to isolate SLOC counts for files which did not require development, but were duplicated for a variety of reasons which could include for configuration management purposes, or were computer generated.

8.7 -trunc

Truncation threshold.

The truncation threshold specifies the maximum number of characters allowed in a logical SLOC. Additional characters will be truncated. The default value is 10,000. If -trunc 0 is specified, no truncation is done. Performance can be significantly degraded if truncation is too high.

8.8 -cf

Support handling ClearCase filenames.

The ClearCase application appends version information to the filename, starting from '@@'. This option requires the UCC tool to handle the original filename instead of the ClearCase-modified filename by stripping off the '@@' and any characters after that and before the extension.

8.9 -dir <dirA> [<dirB>] [filespecs …]

Specify directories containing files to be counted and/or compared.

The -dir switch causes the UCC to look for files to be counted in the specified directory and its subdirectories. If extensions are provided, it will select only files with those extensions. Any default or specified input filelists will be ignored.

If the -d (for differencing) is not present, the UCC will only look for <dirA> and filespecs. If the -d is present, the UCC will look for dirA, dirB and filespecs. The directories are searched recursively for files that match the filespecs.

dirA Name of the top level directory. If the option *-d* is provided, dirA is the directory of *Baseline* A. Otherwise, it specifies the directory to be counted with the counting function. The directory search is recursive so that all subdirectories under the top level directory are searched.

dirB Name of the top level directory of *Baseline* B, used only if the option *-d* is given. The directory search is recursive so that all subdirectories under the top level directory are searched.

filespecs Specifications of file extensions to be counted/compared; wildcard chars ? * are allowed. Use a space between *filespecs*; for example, *.cpp *.c

Examples:

UCC dirA *.*

This command will do counting only, since there is no -d, and will look in the directory dirA recursively to find all files. Any files with extensions recognized by the UCC will be counted

UCC dirA

This command is equivalent to UCC dirA *.*. It will do counting only, and will look in the directory dirA recursively to find all files. Any files with extensions recognized by the UCC will be counted.

UCC dirA *.c *.cpp *.f *.for *.f77 *.f90 *.f03 *.hpf

This command will do counting only, since there is no –d, and will look in the directory dirA recursively, and will process all files found with the extensions specified (.c, .cpp, .f, .for, .f77, .f90, .f03, *.hpf). Notice that the file extensions are separated by a space but have no comma between.

UCC -d dirA *.*

This command is improperly formed, as it specifies -d for differencing, but two directories must be provided.

UCC -d dirA dirB *.*

This command will difference the files found in dirA with the files found in dirB. The *.* direct the UCC to process all files with an extension recognized by the UCC.

UCC -d dirA dirB *.java *.sql

This command will difference files with the extension .java or .sql found in directories dirA and dirB recursively. No other files will be processed.

UCC -d dirA dirB

This command will difference files found in directories dirA and dirB recursively with any extension that UCC recognizes.

8.10 -outdir <dirname>

Prepend the dirname to the output files created.

This command allows the user to specify a directory name to be prepended to the output reports. If the directory does not exist, UCC will create it. The reports will be generated in the specified directory. This allows the user to set up a batch job with multiple UCC runs and not have the output reports overwritten.

Examples:

UCC -outdir test1

This command will look for the default input list fileListA.txt, count the SLOC within the files, and write the output reports into the directory test1.

UCC -outdir test2 -i1 test2files

This command will open the input list of files in test2files, read them in, count the SLOC, and write the output reports into directory test2.

8.11 -extfile <extfilename>

The UCC uses a file's extension to associates files to be counted with language counting modules. One or more extensions may be associated with a given language counter. The table in section 0 lists the current languages and their associated file extensions.

The -extfile switch allows users to modify the file extension mapping by specify which file extensions are to be associated with the language counters. This gives the user more flexibility in determining which files will be counted. It also will be useful as more languages are added to the UCC, as some extensions may be used by more than one language.

Data files will be only be counted for physical SLOC, and only if the users uses an -extfile command to specify the data file extensions. Refer to section 0 for more information.

The file named by <extfile> is a list that associates user-specified extensions with UCC language counters.

The extfile command allows a user to create and specify the name of a file that maps file extensions to UCC language counters. This allows the user to include non-default extensions, or remove default extensions. There are no spaces between the language, the equal sign, or the extensions. The

extensions are separated by a comma. Comments may be placed within square brackets, for example [comment], and may placed anywhere in the extfile.

Each line in the file should have the form:

<language>=.<ext1>[,.<ext2>,…<extn>]

The -extfile option gives the user great flexibility in tailoring specific runs. For instance, suppose a user wishes to count only the files written in Fortran 77 as indicated by having the extension .f77. Placing the Fortran=.f77 command into the extfile would accomplish that. If a user wishes to count Fortran files generated with a non-standard extension, such as .foo. The user can place the command Fortran=.foo in the extfile.

Examples:

Java=.java,.javax

This line would cause any file with the extension .java or .javax to be counted with the java language counter. No other extensions will be counted with the java language counter. All other languages will use the default extensions.

Ada=.ada,.a

Fortran=.f,.for

These lines would cause the Ada counter to count only files with extensions of .ada or .a. All other Ada files with other extensions will be ignored. Also, the Fortran counter will count only files with extensions of .f or .for. All Fortran files with other extensions will be ignored.

8.11.1 File Extension Mapping Names

For reference, the current file extension mappings are shown below. Be sure to use the Internal UCC Language Name in the -extfile mapping. For example, don't use C/C++, use C_CPP.

Language	Internal UCC Language Name (for use in the –extfile option)	File Extensions
Ada	Ada	.ada, .a, .adb, .ads
ASP, ASP.NET	ASP	.asp, .aspx
Bash	Bash	.sh, .ksh
C Shell Script	C-Shell	.csh, .tcsh
C#	C#	.cs
C/C++	C_CPP	.cpp, .c, .h, .hpp, .cc, .hh
ColdFusion	ColdFusion	.cfm, .cfml, .cfc
ColdFusion Script	CFSCRIPT	.cfs
CSS	CSS	.css
Data		Use file mapping with Datafile=<ext> (see below)
Fortran	Fortran	.f, .for, .f77, .f90, .f95, .f03, .hpf
HTML	HTML	.htm, .html, .shtml, .stm, .sht, .oth, .xhtml
Java	Java	.java
JavaScript	JavaScript	.js
JSP	JSP	.jsp
Makefiles	Makefile	.make, .makefile, (files named Makefile)
MATLAB	MATLAB	.m
NeXtMidas	NeXtMidas	.mm
Pascal	Pascal	.pas, .p, .pp, .pa3, .pa4, .pa5
Perl	Perl	.pl, .pm
PhP	PHP	.php
Python	Python	.py
Ruby	Ruby	.rb
SQL	SQL	.sql
VB	Visual_Basic	.vb, .frm, .mod, .cls, .bas
VBScript	VBScript	.vbs
Verilog	Verilog	.v
VHDL	VHDL	.vhd, .vhdl
X-Midas	X-Midas	.txt
XML	XML	.xml

8.11.2 Data file counting

The user must specify Datafile=<ext> to invoke the data counting function. The tool will output only physical counts and only for the extensions specified.

Examples:

Datafile=.dat

Datafile=.dat,.txt

The reason the –extfile must be used for counting data files is that the extension .txt, which is common for data files, was already assigned to the X-Midas counter.

8.12 -unified

Directs the UCC to print counting results to a single unified language file name TOTAL_outfile.csv.

If the -unified command is not specified, a report is produced for each language and is named <Lang>_outfile.csv, which is an Excel format.

8.13 -ascii

Generate reports in text format. The default report format is .csv, which opens directly into Excel.

8.14 -legacy

Generate reports in legacy format.

Use this option to retain the output files formats supported by Difftool, CodeCount Tools Release 2007.07 and earlier versions. The default format has several additional fields.

8.15 -nocomplex

Do not process or report keywords or complexity metrics.

Using this switch will reduce the processing time.

8.16 -nodup

Do not search for duplicate files.

The duplicate file processing is both memory and processor intensive, as the process analyzes each file and compares to potential duplicates character by character. If the user does not need to separate out duplicate files within a baseline, the -nodup switch will disable the duplicate search decreasing processing time significantly. Any duplicates will be reported as unique files.

8.17 –nolinks

Do not follow symbolic links on Unix based systems.

This disables following symbolic links to directories and counting of links to files. This can prevent duplicate file counts on Unix/Linux systems.

9 Performance Issues

Performance is an issue with many facets, among them processor speed, memory size and utilization, bus speed, and program logic. An attempt is made here to point out many of the factors which can affect the performance of running the UCC, along with suggestions for improving the performance.

Certain factors, such as the speed of the CPU, or the amount of memory in the computer, are beyond the scope of this document except to point out to your IT manager that it is time for an upgrade. However, we will attempt to point out cases where performance improvements can be made.

9.1 Compiler Optimization

The UCC is distributed as open source code. Users download the code and compile the executable using any ANSI-standard C++ compiler. Many compilers offer a variety of optimization levels that include speeding up the execution of the program, reducing the size of the program in memory, streamlining low level code that is used multiple times, etc. Typically, you gain speed at the expense of space needed for the resulting executable, or you can get a smaller executable that uses less memory, but doesn't run as fast. The user should consult the user's manual for the compiler being used for the appropriate compiler optimization techniques.

9.1.1 Microsoft Visual Studio

In Microsoft Visual Studio, there are options for General Whole Program Optimization, including Use Link Time Code Generation, Profile Guided Optimization – Instrument, Profile Guided Optimization-Optimize, Profile Guided Optimization-Update, and No Whole Program Optimization. Under the C/C++ Optimization men there are options for Minimize Size (/O1), Maximize Speed (/O2), Full Optimization (/Ox), Custom, Inherit from parent or project defaults, or Disabled (/Od). It is left to the user to experiment with the optimization available with the user's compiler in order to meet their particular situation.

9.1.2 GNU g++

The GNU g++ provides a range of general optimization levels, numbered from 0-3, as well as individual options for tailored optimization. The optimization level is specified on the compilation command line with a -O*LEVEL* switch, where *LEVEL* is a number from 0 to 3. The levels are described below.

-O0 or no -O option (default) – No optimization is performed, and the source code is compiled in the most straightforward way. This is the best option to use when debugging.

-O1 or -O – Common forms of optimization are applied that do not require any speed-space tradeoffs. Executables should be smaller and faster than with -O0. Compilation time can actually be less than when compiling with -O0.

-O2 – Includes the optimizations in -O1, plus some further optimizations. No speed-space tradeoffs are used, so the executable should not increase in size. The compiler will require more memory and time for the compilation, but should not increase the executable size. This is the best option to use when producing a deployable program.

-O3 – Includes more expensive optimizations, such as function inlining, as well as the optimizations of levels -O1 and -O2. This option may increase the speed of the executable, but can also increase the size. Under certain circumstances, it might make the program run slower.

-funroll-loops – This option is independent of the other optimization options. It turns on loop-unrolling, and will increase the size of the executable. It may or may not improve speed.

-Os – This option will produce the smallest possible executable, valuable for systems constrained by memory or disk space. In some cases, a smaller executable will run faster, due to better cache usage.

9.2 UCC Switches Affecting Performance

9.2.1 -trunc

Meaning: Truncation threshold

The truncation threshold specifies the maximum number of characters to be processed in a logical SLOC. Additional characters will be truncated. The default value is 10,000. If -trunc 0 is specified, no truncation is done. Performance can be significantly degraded if truncation is too high. The tradeoff is that when two SLOC are being compared between two programs, the comparison is done character by character, but is stopped when the truncation threshold is met. If there is no difference before the threshold is met, the SLOC are considered unmodified, but if there is a difference after the threshold, the SLOC should have been identified as modified, rather than unmodified. The more long statements (> threshold) the more likely the SLOC identification is to be incorrect. The user must make the trade determination of whether to expend more processing time to process longer statements for greater accuracy, or risk having incorrect SLOC modified/unmodified counts but with a quicker execution speed.

9.2.2 -nodup

Meaning: Do not search for duplicate files.

By default, the UCC looks within each baseline for files whose SLOC are identical to the SLOC in other files. The first file is considered unique, but all the other identical files are considered duplicates. Comments and blank lines are not considered in the duplicate processing; i.e. if the comments and/or blank lines change but the SLOC are still duplicates, the file is still considered a duplicate. If the SLOC have been rearranged but no characters have been modified, they are still considered duplicate.

The duplicate files are counted and reported separately from the unique files for the purpose of measuring work done. The Duplicates-<LANG>_outfile.csv, where <LANG> is the language, will have the counts, and the DuplicatePairs.csv file will identify the original and duplicate file pairs.

The duplicate processing is compute intensive, and increases the execution speed. The user can choose to suppress the duplicate processing by using the -nodup switch on the command line. Then all files will be considered unique and will be included in the standard reports.

9.2.3 -nocomplex

Meaning: Do not produce keyword counts or complexity metrics

By default, the UCC counts keywords and directive, nested loops, and other metrics useful in evaluating complexity. While this is not a major computational task, some processing time can be eliminated by using the -nocomplex switch to suppress the complexity metrics.

9.3 Large Jobs

Large jobs can be defined many ways. A large job may have large amounts of average size files, or a smaller amount of very large files, or files with very long lines of code. Computers with more memory, more disk space, and/or more processor speed will be able to process larger jobs using less time. This section will give strategies on how to do the best you can with what you have.

9.3.1 Memory Limitations

Symptom: the UCC process starts out fine, reporting progress as it goes, and then it hangs. Most likely this is a problem where the process has run out of memory. Current research is being done to perform an analysis at the start of a UCC run to predict if the computer has enough memory to run the task, and notify the user if the prediction is that the job is too large.

9.3.2 Process One or a Few Languages at a Time

If the input files are written in multiple languages, the UCC can be directed to process only certain languages, or even certain extensions within a given language. Refer to the section on the -dir switch and the section on extension file mapping for more information.

9.3.3 Divide and Conquer

The input files can be regrouped into lesser amounts of files. This can be done by creating extra input file lists, giving each a different name, and dividing up the input files into the various lists. See section 3.1.2.1 for tips on creating file lists. For example, the user creates 3 file lists called fileList1.txt, fileList2.txt, and fileList3.txt. Each file list can be run separately using the commands:

UCC -i1 fileList1.txt

UCC -i1 fileList2.txt

UCC -i1 fileList3.txt

Note that the output reports will all be written into the same working directory, and any reports that have the same name will overwrite the previous reports.

9.3.4 Preserving Output Files from Multiple Runs

If multiple UCC runs are to be made consecutively, the user can separate the output report files by using the -outdir command.

For example, the following three runs will output reports into different directories.

UCC -i1 fileList1.txt -outdir fileList1

UCC -i1 fileList2.txt -outdir fileList2

UCC -i1 fileList3.txt -outdir fileList3

The UCC will write the output reports into the directory specified by the -outdir file. If the output directory does not exist, UCC will create it. Consequently the output reports are segregated by run, and are not overwritten. This allows the user to create a batch job or command file to execute multiple UCC runs sequentially where each consecutive run writes the output reports into a different directory.

9.3.5 Difference in Two Steps

When differencing two large baselines, the user can count each baseline separately to get the duplicate file information. Then difference the two baselines with duplicate processing turned off. See the -nodup switch for more information.

9.3.6 Long Line Truncation

If there are several long lines (greater than 10,000 characters), as indicated by entries in the error log, the user can set the truncation to a smaller number using the -trunc switch. Any differences that occur after the truncation will not be detected.

10 Language Specific Information

10.1 Ruby

The Ruby programming language allows continuation characters of "," and ".", but also allows a default continuation if the syntax of a line is not complete. The compiler assumes that the next line is a continuation line. UCC does not examine the syntax. Consequently, if a line is a continuation of the previous line due to context specific information, UCC will not understand that and will count it as a separate line. This could impact the LSLOC count, making it higher than it should be. The PSLOC count will not be affected.

10.2 Fortran

Currently Fortran version 77 and lower uses a fixed format where any character (except 0) in column 6 is a continuation character indicating that this line is a continuation of the previous line. Later versions (F90 and above) do not have this restriction; the continuation character is an & at the end of the line. Currently the UCC is using the format for F90 for all Fortran code. There are plans to develop a separate counter for Fortran 77 and lower.

10.3 JavaScript

The JavaScript counter does not count the statement that is not terminated by a semicolon.

11 References

[1] R.E. Park, "Software Size Measurement: A Framework for Counting Source Statements", Technical Report CMU/SEI-92-TR-20 ESC-TR-92-020, 1992.

[2] B. Boehm, C. Abts, S. Chulani, "Software development cost estimation approaches: A survey", *Annals of Software Engineering, 2000*.