



THE  
POWER  
TO KNOW.

# **SAS<sup>®</sup> Concept Creation for SAS Text Miner 5.1 User's Guide**



The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2011.  
*SAS® Concept Creation for SAS Text Miner 5.1: User's Guide*. Cary, NC: SAS Institute Inc.

### **SAS® Concept Creation for SAS Text Miner 5.1: User's Guide**

Copyright © 2011, SAS Institute Inc., Cary, NC, USA

All rights reserved. Produced in the United States of America.

**For a hard-copy book:** No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, or otherwise, without the prior written permission of the publisher, SAS Institute Inc.

**For a Web download or e-book:** Your use of this publication shall be governed by the terms established by the vendor at the time you acquire this publication.

The scanning, uploading, and distribution of this book via the Internet or any other means without the permission of the publisher is illegal and punishable by law. Please purchase only authorized electronic editions and do not participate in or encourage electronic piracy of copyrighted materials. Your support of others' rights is appreciated.

**U.S. Government Restricted Rights Notice:** Use, duplication, or disclosure of this software and related documentation by the U.S. government is subject to the Agreement with SAS Institute and the restrictions set forth in FAR 52.227-19, Commercial Computer Software-Restricted Rights (June 1987).

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

1st electronic book, May 2011

SAS® Publishing provides a complete selection of books and electronic products to help customers use SAS software to its fullest potential. For more information about our e-books, e-learning products, CDs, and hard-copy books, visit the SAS Publishing Web site at [support.sas.com/publishing](http://support.sas.com/publishing) or call 1-800-727-3228.

SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.

---

# Contents

---

<b>About This Book .....</b>	<b>ix</b>
Audience .....	ix
Prerequisites .....	ix
Conventions .....	ix
<b>1 About SAS Concept Creation for SAS Text Miner .....</b>	<b>1</b>
1.1 What is SAS Concept Creation for SAS Text Miner? .....	1
1.2 How Do Concepts Work in SAS Text Miner? .....	2
1.3 Benefits to Using SAS Concept Creation for SAS Text Miner .....	2
1.4 How Does SAS Concept Creation Work with SAS Text Miner? .....	4
1.5 The Architecture .....	4
<b>2 Using the Interface .....</b>	<b>5</b>
2.1 Your First Look at SAS Concept Creation for SAS Text Miner .....	5
2.2 The SAS Concept Creation Menus .....	7
2.2.1 About the Availability of Menus and Menu Selections .....	7
2.2.2 About Menus .....	7
2.2.3 The File Menu .....	7
2.2.4 The Edit Menu .....	8
2.2.5 The View Menu .....	9
2.2.6 The Build Menu .....	9
2.2.7 The Project Menu .....	9
2.2.8 The Concept Menu .....	10
2.2.9 The Testing Menu .....	11
2.2.10 The Document Menu .....	11
2.3 The Status Bar .....	13
2.4 The Standard Toolbar .....	13
2.5 The Taxonomy Tab .....	14
2.6 The Right Window Tabs .....	16
2.6.1 Overview of the Tabs .....	16
2.6.2 The Definition Tab .....	17
2.6.3 The Testing Tab .....	18
2.6.4 The Data Tab .....	19
2.6.5 About the Document and Concordance Tabs .....	21

---

2.6.5.A Overview of the Document Tab .....	21
2.6.5.B The Document Tab .....	21
2.6.5.C The Document Tab as Concordance .....	22
2.6.5.D The Document Tab as Browser Interface .....	23
2.6.5.E The Components of the Document Tab .....	24
2.7 The Project Settings Windows .....	25
2.7.1 Project Settings Overview .....	25
2.7.2 The LITI Tab .....	26
2.7.3 The Misc Tab .....	27
2.7.4 The Concordance Tab .....	29
2.8 The Miscellaneous Windows .....	31
2.8.1 The Select a Language Window .....	31
2.8.2 The Enter Names Window for UTF-8 Languages .....	33
2.8.3 The Number of Taxonomy Nodes Window .....	34
2.8.4 The Concordance Windows That Are Available through the Testing Tab .....	35
2.8.5 The Tree Find Window .....	36
2.8.6 The Tree Replace Window .....	38
2.8.7 The Text Find and Replace Windows .....	39
2.8.8 The Compile Concepts Tab .....	39
2.8.9 The Concordance Windows .....	40
2.8.10 The Syntax Check Window .....	41
2.8.11 The Best Matches Window .....	42
2.8.12 The Concept Priorities Window .....	43
2.8.13 SAS Concept Creation Status Window Example .....	45
2.9 The Drop-down Taxonomy Node Operations .....	45
2.9.1 The Project Name Node Operations .....	45
2.9.2 The Language Node Operations .....	46
2.9.3 The Concepts Node Operations .....	47
2.9.4 The Individual Concept Node Operations .....	48
<b>3 Creating Projects .....</b>	<b>49</b>
3.1 Overview of Creating Projects .....	49
3.2 Start SAS Concept Creation .....	50
3.3 Create a New Project .....	51
3.4 Saving the Project .....	56
3.4.1 Overview of the Save Operation .....	56
3.4.2 Manually Save an Existing Project .....	56
3.4.3 Save a Duplicate Project .....	56

---

3.5 Access an Existing Project .....	57
3.6 Choosing Project Settings .....	59
3.6.1 Overview of Project Settings .....	59
3.6.2 Choose How Matches Are Returned .....	59
3.6.3 Choose the Concordance Operations .....	61
3.6.4 Choose Miscellaneous Operations .....	62
3.7 Navigating through the Taxonomy .....	63
3.8 Specify the .li File in SAS Text Miner .....	64
<b>4 Writing Concept Definitions .....</b>	<b>65</b>
4.1 Overview of Definitions .....	65
4.2 Before You Write Your Definitions .....	67
4.3 The Rule Types .....	68
4.4 The Building Blocks .....	69
4.4.1 Overview of the Building Blocks .....	69
4.4.2 Case-Insensitive Matching .....	69
4.4.3 Entering Comments into Rules .....	69
4.4.4 The Tokens .....	70
4.4.5 The _c Marker .....	70
4.4.6 The _w Term .....	70
4.4.7 The _cap Term .....	71
4.4.8 The > Symbol .....	71
4.4.9 The Quotation Marks .....	72
4.4.10 The Parentheses, Square Braces, and Curly Braces .....	72
4.4.11 The Commas .....	72
4.4.12 The Colons .....	73
4.4.13 The Spaces .....	73
4.4.14 The Part-of-Speech Tags .....	74
4.4.15 The Export Feature .....	74
4.4.16 The Regular Expressions .....	75
4.4.17 The Priorities and Project Settings .....	75
4.4.17.A Overview of Priorities .....	75
4.4.17.B Choose Project Settings .....	76
4.4.17.C Seeing the Priorities for the Taxonomy .....	78
4.5 The Operators .....	79
4.5.1 The Boolean Operators .....	79
4.5.1.A The ALIGNED Operator .....	80
4.5.1.B The AND Operator .....	80
4.5.1.C The OR Operator .....	80

---

4.5.1.D The DIST_n Operator .....	80
4.5.1.E The ORDDIST_n Operator .....	81
4.5.1.F The SENT Operator .....	81
4.5.1.G The SENT_n Operator .....	81
4.5.1.H The SENTSTART_n Operator .....	81
4.5.1.I The SENTEND_n Operator .....	82
4.5.2 The Stemming Operator .....	82
4.5.3 The PARA Operator .....	82
4.5.4 The Operators for Coreference Resolution .....	83
4.6 Some Rule Examples .....	83
4.6.1 The Classifier Rules .....	83
4.6.2 The Sequence of Classifier Entries .....	84
4.6.3 Context Matching .....	85
4.6.4 Match within Context .....	86
4.6.5 Eliminating Partial Matches .....	88
4.6.6 Disambiguating Concepts .....	89
4.6.7 Exporting Classifiers .....	91
4.6.8 Setting Priorities for Overlapping Matches .....	94
4.6.9 The Part-of-Speech Tags in a Definition .....	96
4.6.10 The Regular Expressions in a Definition .....	97
4.6.11 The Sentence Operator in a Definition .....	98
4.6.12 The Paragraph Operator in a Definition .....	100
4.6.13 The DIST Operator in a Definition .....	102
4.6.14 The ORDDIST Operator in a Definition .....	104
4.7 Locating Facts .....	107
4.7.1 Overview of Facts .....	107
4.7.2 The Predicate Sequence Example .....	107
4.7.3 The Predicate Examples .....	109
4.8 The Coreference Operators .....	114
4.8.1 Overview of Coreference .....	114
4.8.2 How to Use the Coreference Operator .....	114
4.8.3 How to Use the _ref Operator with the > Symbol .....	115
4.8.4 How to Use the _ref Operator with the Forward or Backward Symbols .....	116
4.8.4.A Limiting Matches to Those That Follow or Precede a Coreference Match .....	116
4.8.4.B Matching with the Forward Symbol .....	116
4.8.4.C Matching with the Preceding Symbol .....	117
4.8.5 Coreference in a Classifier Definition Example .....	118

---

4.8.6 Assigning New Concept Names to Coreference Matches .....	118
4.8.7 Rank Coreference Definitions and Eliminate False Positives .....	119
4.9 XML Fields in Rules .....	120
4.9.1 Overview of XML Field Matching .....	120
4.9.2 The SEQUENCE Rule with an XML Field Example .....	121
4.9.3 Matching More than One XML Field .....	122
4.10 Writing Multiple Rules for One Definition .....	123
4.11 Troubleshooting Your Rules .....	123
<b>Part 2: Testing .....</b>	<b>125</b>
<b>5 Assembling Testing Sets .....</b>	<b>127</b>
5.1 Overview of Assembling Testing Sets .....	127
5.2 Creating Testing Folders .....	128
5.2.1 Create a Testing Directory While You Set Paths .....	128
5.2.2 Create and Set a Path to the Central Repository .....	134
5.2.3 Manually Create a Testing Folder and Set a Path for a Newly Created Concept .....	137
5.3 Collecting Test Files .....	137
5.4 Import Test Files .....	139
5.5 Delete Testing Files .....	142
<b>6 Testing the Concept Definitions .....</b>	<b>143</b>
6.1 Overview of Testing .....	143
6.1.1 Windows .....	144
6.2 Using the Testing Window .....	144
6.2.1 The Testing Window Messages .....	144
6.3 Batch Testing .....	146
6.3.1 Overview of Batch Testing .....	146
6.3.2 Option 1A: Batch Testing All of the Documents for One Concept .	147
6.3.3 Option 1B: Batch Testing the Testing Taxonomy or Out-of-Concept Files .....	148
6.4 Testing with the Document Window .....	150
6.4.1 Overview of Document Window Operations .....	150
6.4.2 Test Using the Document Window .....	151
6.4.3 Testing a Web Page in the Document Window .....	152
6.4.4 Using Windows Commands .....	154
6.4.5 Copy and Paste a Test File .....	155
6.4.6 Using Clear Test Document .....	155
6.4.7 Refreshing the Taxonomy Tree .....	156

---

6.4.8 Changing the Font Size of a Tested Document .....	156
6.4.9 Removing Markup Tags .....	157
6.5 Test a Central Repository .....	157
6.6 Comparing Test Results .....	158
6.7 Import Failing Documents .....	159
6.8 Testing with the Concordance .....	161
6.8.1 An Overview of the Concordance .....	161
6.8.2 The Concordance for the Testing Window .....	161
6.8.3 The Concordance for the Document Window .....	162
6.8.3.A An Overview of the Concordance for the Document Window .....	162
6.8.3.B Determine How the Concordance Is Displayed .....	163
6.8.3.C See the Concordance Terms for a Selected Concept .....	164
6.8.3.D Use the Best Matches Window for All Concepts .....	165
6.8.3.E See the Concordance Terms for All Concepts .....	166
<b>Appendixes .....</b>	<b>169</b>
<b>A Regex Syntax and Part-of-Speech .....</b>	<b>171</b>
A.1 Regular Expressions .....	171
A.1.1 Rules and Restrictions .....	171
A.1.2 Special Characters .....	173
A.1.3 Special Cases .....	174
A.2 Part-of-Speech Table .....	174
<b>B Glossary .....</b>	<b>181</b>
<b>C Recommended Reading .....</b>	<b>185</b>
<b>Index .....</b>	<b>187</b>

---

# About This Book

---

## Audience

SAS Concept Creation for SAS Text Miner (SAS Concept Creation) is designed for subject matter experts who write the complex rules for concepts. These concepts identify the context-sensitive data that exists in your organization's input documents.

## Prerequisites

Here are the prerequisites for using SAS Concept Creation:

- SAS Text Miner loaded onto your machine
- SAS Concept Creation loaded on the same machine as SAS Text Miner, or on a different machine
- License file for SAS Text Miner
- Representative documents where you want to locate metadata

## Conventions

This manual uses the following typographical conventions:

Convention	Description
TGM_ROOT	The root directory where SAS Concept Creation for SAS Text Miner is installed, typically the following: <b>Windows:</b> C:/Program Files/SAS/SAS Concept Creation
.li	The code examples for the .li file are shown in a fixed-width font.
<b>TEST</b> button	The labels for user interface controls are shown in a bold, sans-serif font.

---

Convention	Description
Top	The names of taxonomy nodes appear in a fixed-width font.
<a href="http://www.sas.com">www.sas.com</a>	The hypertext links are shown in a light blue, fixed-width font, and are underlined.

---

# 1

## About SAS Concept Creation for SAS Text Miner

---

- *What is SAS Concept Creation for SAS Text Miner?*
- *How Do Concepts Work in SAS Text Miner?*
- *Benefits to Using SAS Concept Creation for SAS Text Miner*
- *How Does SAS Concept Creation Work with SAS Text Miner?*
- *The Architecture*

### 1.1 What is SAS Concept Creation for SAS Text Miner?

In most organizations it is necessary to identify metadata, or data on information. This metadata is located in your documents, created internally and externally, and stored in your company's repositories.

SAS Concept Creation for SAS Text Miner (SAS Concept Creation) is an add-on product that works with SAS Text Miner. SAS Text Miner uses the binary (.1i) files created in SAS Concept Creation to locate custom entities in documents that are input to SAS Text Miner.

Using the intuitive, Windows interface in the SAS Concept Creation application, subject matter experts write complex rules to define each concept in the taxonomy. This taxonomy is output as a .1i file. To use the .1i file that SAS Concept Creation generated, you set properties in the **Text Parsing** node of SAS Text Miner.

---

---

## 1.2 How Do Concepts Work in SAS Text Miner?

*Concepts* is another word for the term *custom entities* that is used in SAS Text Miner. SAS Text Miner identifies standard entities such as Percent, Phone, Time, and so on. In order to identify custom entities, SAS Text Miner uses the concepts that you define in SAS Concept Creation. Use the binary files created in SAS Concept Creation to locate custom entities in input documents using SAS Text Miner.

The word *entity* is used in SAS Text Miner to refer to the predefined metadata that can be extracted from unstructured text. Use SAS Concept Creation for SAS Text Miner to specify the *concepts* that locate the metadata that you seek according to the definitions that you write. Concepts are *custom entities* that are similar to entities. However, concepts enable you to specify the semantic relationships between terms that improve the accuracy of entity matching.

## 1.3 Benefits to Using SAS Concept Creation for SAS Text Miner

SAS Concept Creation expands the benefits available in SAS Text Miner:

Context sensitive matching

Limit concept matching to those matches that occur within the specified context. For example, match New York but not New York City.

Syntax building blocks

Write your definitions to locate concept matches using parts of speech, logical operators, regular expressions, and separator characters.

Concept disambiguation

Return only the specific concept that you are seeking. For example, differentiate between Giants football and Giants baseball.

Relational concepts

Return related concepts. For example, locate the string *Drew Faust is president of Harvard University*, where *Drew Faust* and *Harvard University* are concepts.

---

### Fact extraction

Extract facts from seemingly unrelated pieces of data, similar to relational concepts. Specify operators between the concepts that together form a fact to return the entire string. For example, match Tide is produced by Procter & Gamble.

### Stemming

Locate matches on all, or only the noun or verb, forms of a word.

### Multiple types of definitions

There are a few types of definitions that you can use to locate matches.

### Write multiple rules for one definition

Match on any rule, within a concept definition, in an input document and return a match on this concept.

### Write different types of rules for one definition

Specify different types of rules for each concept definition.

### Determine how SAS Concept Creation treats overlapping, identical, or duplicate matches

Specify the appropriate settings using the Project Settings - LITI dialog box to determine the matching process in these cases.

### Coreference operators

Use coreference operators to write rules that return the canonical form of a word along with the referring term.

### Apply concepts as custom entities

After you develop and test the taxonomy, specify the path to the .li file in the **Text Parsing** node of SAS Text Miner. Concept definitions are applied as custom entities to incoming documents.

---

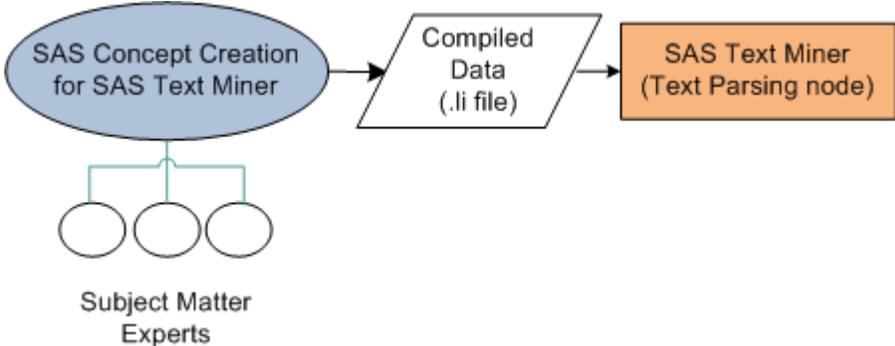
# 1.4 How Does SAS Concept Creation Work with SAS Text Miner?

The functionalities of SAS Concept Creation are fully integrated into the SAS Text Miner user interface. Anyone can use the SAS Concept Creation interface to develop taxonomies, define concepts, and write definitions for these concepts. In addition, SAS Concept Creation enables you to test your concepts to see how well their definitions perform using selected, or real-world, testing documents. After you develop a project, the concepts and their definitions are saved into a .li file that you can import and use with SAS Text Miner.

# 1.5 The Architecture

Use the architecture diagram below to gain an overview of the project development processes.

*Figure 1-1 SAS Concept Creation for SAS Text Miner Architecture*



---

# Chapter: 2

## Using the Interface

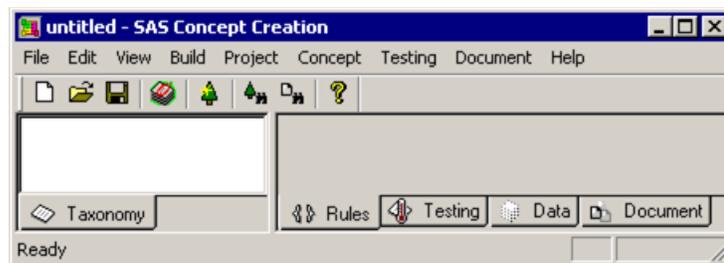
---

- *Your First Look at SAS Concept Creation for SAS Text Miner*
- *The SAS Concept Creation Menus*
- *The Status Bar*
- *The Standard Toolbar*
- *The Taxonomy Tab*
- *The Right Window Tabs*
- *The Project Settings Windows*
- *The Miscellaneous Windows*
- *The Drop-down Taxonomy Node Operations*

### 2.1 Your First Look at SAS Concept Creation for SAS Text Miner

To access the SAS Concept Creation for SAS Text Miner (SAS Concept Creation) user interface, go to **Start** —> **Programs** —> **SAS Concept Creation** —> **SAS Concept Creation**.

*Display 2-1 Main Window*



---

The components of the main window are listed below from top to bottom:

**Program and Project title bar**

display the name of the program and the title of the current project. (The title only appears after you create a new project.)

**Menu bar**

access drop-down lists for project tasks. For more information, see Section 2.2 *The SAS Concept Creation Menus* on page 7.

**Standard toolbar**

click shortcut buttons for some operations. For more information, see Section 2.4 *The Standard Toolbar* on page 13.

**Taxonomy tab**

create, edit, and see the hierarchical structure of the concepts that define your project. For more information, see Section 2.5 *The Taxonomy Tab* on page 14.

**Definition tab**

(**Rules** tab changes to **Definition** tab when you work in concepts area of taxonomy) write the definitions that classify input documents into concepts. For more information, see Section 2.6.2 *The Definition Tab* on page 17.

**Testing tab**

test your definitions against the testing sets of documents that you assemble. For more information, see Section 2.6.3 *The Testing Tab* on page 18.

**Data tab**

specify the priorities, case sensitivity, and the paths to testing documents here. For more information, see Section 2.6.4 *The Data Tab* on page 19.

**Document tab**

see the matches for the tested concept in a single tested document. Also access the concordance operations through this window. For more information, see Section 2.6.5 *About the Document and Concordance Tabs* on page 21.

---

## 2.2 The SAS Concept Creation Menus

### 2.2.1 About the Availability of Menus and Menu Selections

All of the following conditions influence whether a menu or menu selection is available to use:

- Your location in the SAS Concept Creation application. For example, some tasks are available only if you select a tab.
- Whether, or not, you created a project.
- The selections that you choose.

### 2.2.2 About Menus

Menus contain operations that apply to the entire project, or to the currently displayed tab. For example, create a new project, access an existing project, or build a project.

### 2.2.3 The File Menu

Here are the operations that are available in the **File** menu:

**New Project**

access the New Project window where you name, set the path, and choose a language, for your new project.

**Open Project**

locate and access an existing project using the Open window that appears.

**Save Project**

preserve the current project.

**Save Project As**

save the current project and rename a new, duplicate project.

**Exit**

close SAS Concept Creation.

---

## 2.2.4 The Edit Menu

The standard **Undo**, **Redo**, **Cut**, and **Copy** Window commands are located here. The following operations are also included in this menu:

### **Cut All Selections**

use the SHIFT key to select multiple values, such as several taxonomy nodes. To select noncontiguous values, press the CTRL key and select the specific nodes that you want to delete.

### **Copy All Selections**

copy all of the selected nodes. You can paste these nodes into a different area of the taxonomy as duplicates of the existing nodes.

---

**Note:** The **Cut All Selections** and **Copy All Selections** operations delete and copy children, as well as parent, nodes.

---

### **Paste**

paste a single node into your taxonomy. If you select a parent node, all of the children (subnodes) of the selected parent are pasted into the taxonomy. See the related operation **Paste Single Node** below.

### **Paste Single Node**

paste *one* copied node into the taxonomy, as a child of the selected parent node.

### **Text Find**

locate text in the **Document** tab.

### **Text Replace**

access the Replace window that you use to substitute text in the **Document** tab.

### **Tree Find**

use the Find window that appears to search the **Taxonomy** tab for concepts.

### **Tree Replace**

enter text into the Replace window to locate and replace in the **Taxonomy** tab.

---

**Find in All Rules**

search for a matching string in the concept definitions in the Find in All Rules window that appears.

## 2.2.5 The View Menu

Use these commands to hide, or show, the standard **Toolbar** and **Status Bar**. You can also access the following commands:

**Refresh Tree**

update the directory tree in the **Taxonomy** tab when you remove testing messages.

**Taxonomy as Text**

see the taxonomy in text format.

**Number of Taxonomy Nodes**

see a list of the taxonomy nodes and a count of the subnodes in the Number of Taxonomy Nodes window that appears.

## 2.2.6 The Build Menu

The following commands are located in this drop-down menu:

**Compile Concepts**

build a .li file. The **Compile Concepts** tab appears at the bottom of the SAS Content Categorization Studio interface where you can see the results of this operation.

**Abort Compiling Concepts**

stop the process of compiling the concepts. This operation can be used with large concepts projects. When large concepts projects are built, the process of compilation can be lengthy.

## 2.2.7 The Project Menu

The following commands are located in this drop-down menu:

**Add Language**

---

enable the project to be built in a language that you purchased. When you select this operation, the Select a Language window appears. This window contains a drop-down list of the languages that you purchased.

**Delete Language**

select this operation and a SAS Concept Creation status window appears. You can remove the language applicable to the selected taxonomy node.

---

**Note:** If you click **Yes** in the SAS Concept Creation window, you lose all of the nodes and branches that use this language.

---

**Enable Concepts**

enable concept extraction in this project.

**Remove Concepts**

select the language node in the **Taxonomy** tab and choose this operation to delete all of the concepts in the taxonomy.

**Settings**

specify project-wide settings using the Project Settings window that appears.

## 2.2.8 The Concept Menu

The following commands are located in this drop-down menu:

**Add Concept**

add a child node to the parent node that you selected.

---

**Note:** You can specify duplicate concept names, but only if their case is different.

---

**Delete Concept**

remove a node.

**Delete All Selected Concepts**

remove all of the selected concepts.

**Rename Concept**

---

enter the new name of the selected concept.

**Priorities**

access the Concept Priorities window that displays the priority setting for each concept. This setting is specified in the **Priority** field of the **Data** tab.

**Create Directory Tree**

impose a directory structure from the disk to your project or from the project to disk.

## 2.2.9 The Testing Menu

The following testing operations are located in this drop-down menu:

**Import Test Files**

display the names of your test documents in the **Testing** tab.

**Import Failing Test Files**

bring test documents that could, but should not, pass the test for the selected node into the Testing window to test them. For example, you might want to ensure that the term *server* that applies to a restaurant concept does not match a computer concept.

**Delete Selected Test File**

remove the test file that you selected from the **Testing** tab.

## 2.2.10 The Document Menu

The following operations are located in this drop-down menu:

**Clear Test Document**

remove the contents of the **Document** tab.

**Open Test Document**

access a test document in the **Document** tab.

**Save Test Document**

perform a Save operation. This operation copies the testing document that appears in the **Document** tab into the folder of your choice.

**Save Test Document As**

---

perform a Save as operation. Save the changes in a testing document, shown in the **Document** tab, into the directory of your choice.

**Decrease Font Size**

minimize the size of the font for the displayed test file.

**Increase Font Size**

enlarge the size of the font for the displayed test file.

**Remove Tags**

remove any markup language from the testing document.

**Browser**

use this operation and its suboptions with a Web document in the **Document** tab. There are selections that are related to the use of the **Browser** selection:

**Forward**

jump to the next page.

**Back**

return to the previous page.

**Refresh**

update the current Web page.

**Stop**

stop loading the current page.

**Home**

return to the first page that was loaded into the browser.

---

## 2.3 The Status Bar

The **Status Bar** is the horizontal area at the bottom of the SAS Concept Creation interface that indicates the status of the operation that is currently running.

*Display 2-1 Status Bar*



Select **View** → **Status Bar** to hide, or show, the status bar.

## 2.4 The Standard Toolbar

Use the standard toolbar, located below the menu bar, to access some operations. These standard toolbar icons are shortcuts to some, but not all, of the commands available from the menu bar.



Select or deselect the standard toolbar to hide or show the **Toolbar** operation in the **View** menu.

Table 2-1: Standard Toolbar Buttons

Icon	Command
	Click <b>New</b> and the New Project window appears. Name the project, choose a path, and a language for the new project.
	Click <b>Open</b> and the Choose a project file window appears where you locate an existing project file (. tk2).
	Click <b>Save</b> to preserve the changes to the project.

Table 2-1: Standard Toolbar Buttons (Continued)

Icon	Command
	Click <b>Compile Concepts</b> to build the <code>.concepts</code> file.
	Click <b>Refresh Tree</b> to clear the testing messages from the taxonomy tree.
	Click <b>Tree Find</b> to access the Tree Find window to search the taxonomy.
	Click <b>Text Find</b> and the Text Find window appears where you can enter the text that you want to locate.
	Click the question mark icon to access the <i>SAS Concept Creation for SAS Text Miner: User's Guide</i> .

## 2.5 The Taxonomy Tab

By default, the **Taxonomy** tab is displayed when you start SAS Concept Creation. Use this window to see the taxonomy of concepts that you define. If you build your taxonomy with more than one language, an additional language branch is added.

---

Display 2-2 One Project



The following nodes appear in the taxonomy:

LITI\_Demo

see the name of the project.

English

see the language branch.

Concepts

see the unchangeable node name for the concepts extractor.

Top

see the unchangeable node name for the root of each taxonomy.

---

**Notes:** Some of the nodes that are listed above appear only after the related functionality is added to the project. Some of the command shortcuts that are available on the menu and standard toolbars, are also accessible when you right-click on a node in the **Taxonomy** tab.

---

---

## 2.6 The Right Window Tabs

### 2.6.1 Overview of the Tabs

Use the tabs that are located on the bottom right-hand side of the user interface to write definitions, enter data, test the taxonomy, and so on.

*Display 2-3 Concept Tabs*



The table below describes the components of these tabs:

Table 2-2: Window Tab Commands

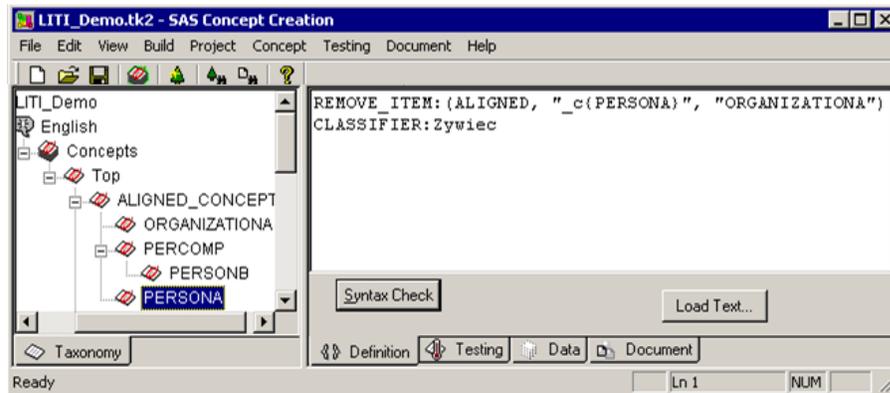
Tab	Purpose
Definition	Write concept rules.
Testing	Test documents against the definitions. Also specify the <b>Concordance</b> operations using selections that are available in this window.
Data	Enter priority and case sensitivity information. Also use this window to specify testing data such as the path to the testing directory.
Document	See the testing results for one document. The Document window becomes the concordance window when the <b>Concordance</b> check box is selected with either <b>Selected concept</b> or <b>All concepts</b> .

---

## 2.6.2 The Definition Tab

Use the **Definition** tab to specify the rule for the selected concept.

*Display 2-4 Definition Tab*



Use the buttons in the **Definition** tab when you define your concepts:

### Syntax Check

access the Concept Syntax Check window, where you check the syntax of your definitions.

### Load Text

load the full text of a file into the **Definition** tab as your concept definition. For example, write a complex definition using a .txt document. Click **Load Text** to access the Open window where you can locate the definition text that you want to load into the **Definition** tab.

### Ln

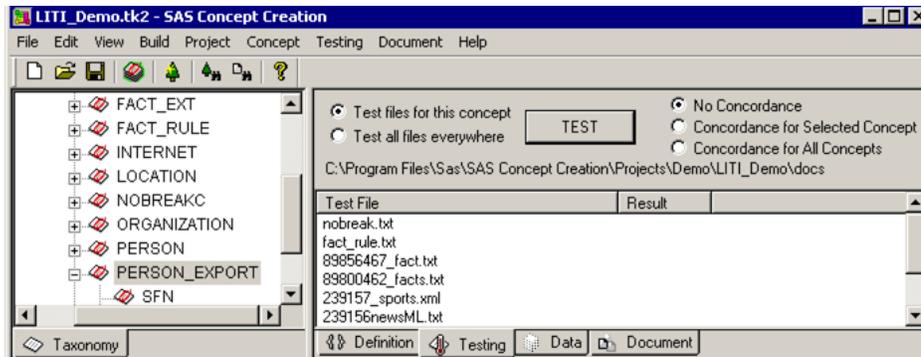
see the line number where your cursor is located. For example, your cursor might appear on Ln 56.

---

## 2.6.3 The Testing Tab

Use the **Testing** tab to check the accuracy of a concept definition against a set of testing documents.

*Display 2-5 Testing Tab*



The operations that are available in the **Testing** tab are explained below:

### **Test files for this concept**

test only the test files that are mapped to this concept in the **Data** tab.

### **Test all files everywhere**

test all of the files in the testing repository against the selected concept. This operation expands the testing process to simulate real-time results.

### **TEST**

start the testing process and SAS Concept Creation displays the results in the **Testing** window.

### **No concordance**

(default) do not perform any concordance operations.

### **Concordance for Selected Concept**

display the terms that match the selected concept in the input document in the concordance window.

### **Concordance for All Concepts**

display the matched concepts for all of the terms in your definitions. These terms appear in the concordance window, with the names of the concepts that they match.

---

The path to the testing file is displayed below these operations and above the following headings:

**Test File**

see a list of the names of all of the test files below this heading. (This list appears after you specify the path to the testing directory in the Data window.)

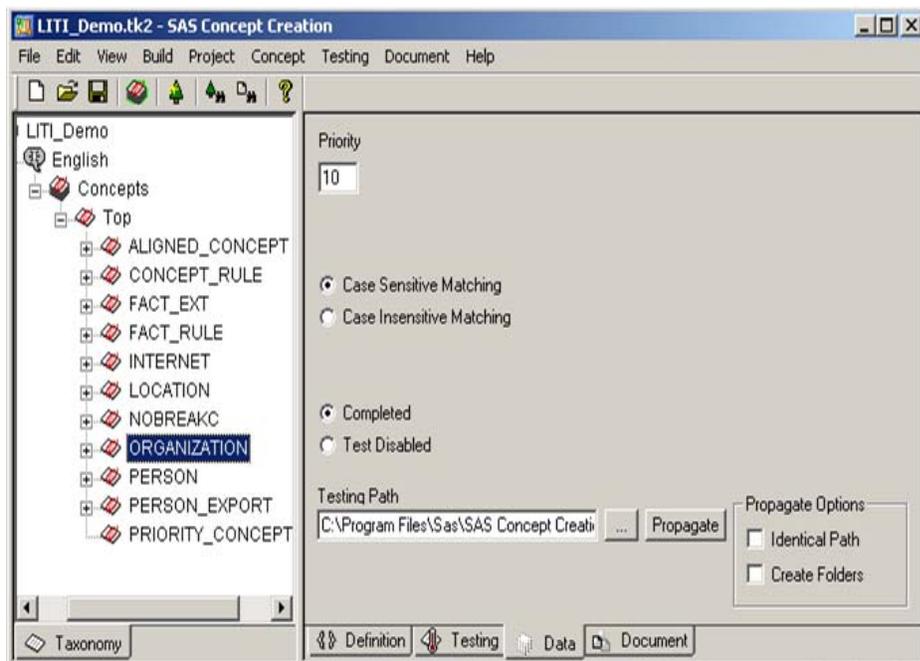
**Result**

display the number of matches for this concept definition.

## 2.6.4 The Data Tab

Use the **Data** tab to enter metadata, the testing path, and other information for each concept.

Figure 2-2 Data Tab



The following table describes the components of the Data window:

---

**Priority**

(default: 10) determine the matching concept when a document matches more than one concept and no other determiner makes one match better than another.

**Completed**

(default) flag this node as finished.

**Test Disabled**

define helper concepts that are evaluated but not exposed to the user.

**Case Sensitive Matching**

(default) match a string in an input document that is an exact, case-sensitive match for the specified text.

**Case Insensitive Matching**

locate a match on a string in an input document when the text of the string is a match, regardless of the case specified by the concept.

**Testing Path**

specify the pathname to the directory that contains the testing documents that are used to analyze this accuracy of this concept definition.

**Propagate**

specify the testing path.

**Propagate Options**

specify either, or both, of the operations under this heading:

**Identical Path**

specify testing paths to the same repository of testing documents.

**Create Folders**

automatically create folders for all of the child concepts.

---

## 2.6.5 About the Document and Concordance Tabs

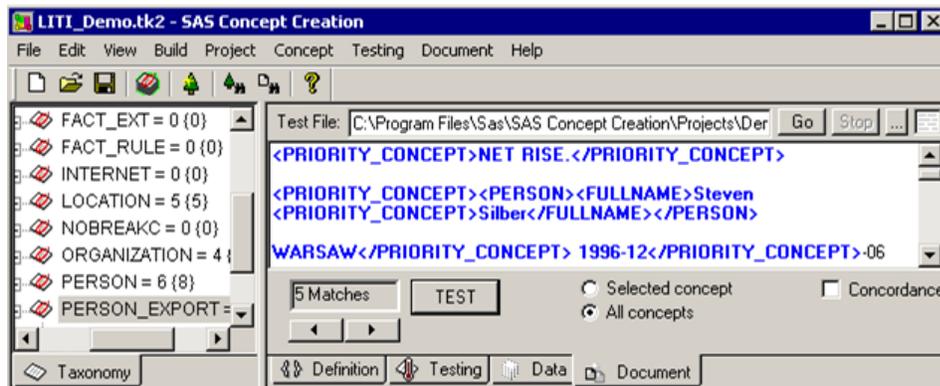
### 2.6.5.A Overview of the Document Tab

The **Document** tab is used to test the text of a test document, a Web page, or any text that you type, or copy and paste, into this window. There is a 1 MB limit for text that is typed, or copied and pasted, into this window. The Document window becomes a concordance window when you choose to use one of the concordance operations.

### 2.6.5.B The Document Tab

The **Document** tab displays the matching concept rule terms and includes the concordance view.

Figure 2-3 Document Tab for Concepts



Select one of the following test operations in the **Document** tab and see the results in the input document:

#### **Selected concept**

(default) test the text in the **Document** tab against the selected concept.

#### **All concepts**

test the selected document against all of the concepts in the taxonomy.

After you select one of the operations listed above, click **TEST** and see the following test results for the selected document:

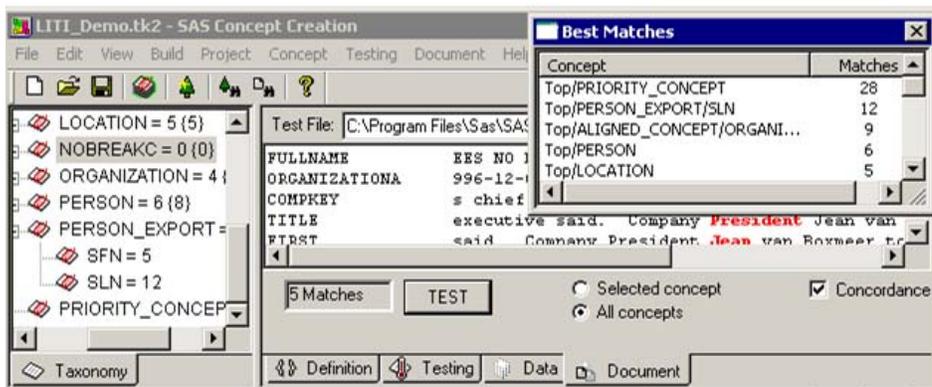
- The matching terms are highlighted in red for the selected concept and blue for all other concepts.
- The PASS or FAIL result is displayed in the testing field to the left of the **TEST** button.
- Jump to the following or preceding match when you click either the  or  arrow.

To use the concordance selection, see Section 2.6.5.C *The Document Tab as Concordance* below.

### 2.6.5.C The Document Tab as Concordance

A concordance is an ordered list of matched terms for the selected concept. You specify this ordering in the Project Settings - Concordance window. The concordance view displays only the terms that match the concept definition in the **Document** tab according to the specifications that you set in the Project Settings - Concordance window. You can also see the results for all concepts in the Best Matches window.

*Display 2-6 Concordance View*



By default, the concordance does not apply. Choose from the following combinations of selections in the Document window to use the concordance:

#### **Concordance and Selected Concept**

display the terms that match the selected concept in the input document in the concordance window that appears.

---

## Concordance and All Concepts

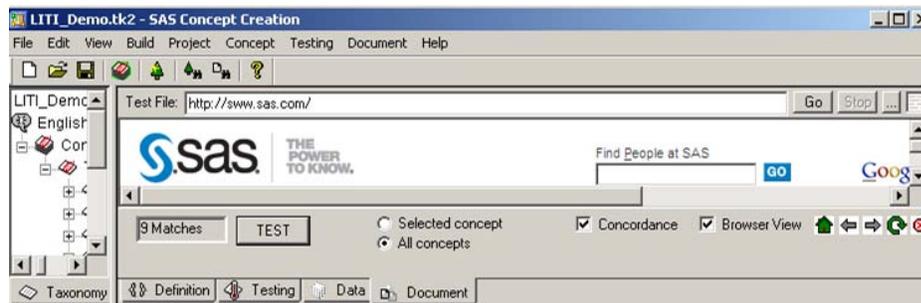
display the matched concepts for all of the terms in your definitions. These terms appear in the concordance window, with the names of the concepts that they match.

After you click **TEST**, the **Document** tab changes into the concordance view. Definition matches appear in list format. If you select **All concepts**, the Best Matches window displays a list of matching concepts with the total number of their matches. For more information, see Section 2.8.11 *The Best Matches Window* on page 42.

### 2.6.5.D The Document Tab as Browser Interface

The **Document** tab can also be used as a Web browser to test Web documents.

*Display 2-7 Document Window as Web Browser*



To test a Web document, select **Browser View**. When the results appear, you can also use the Best Matches window to see the total count of the matches. For more information, see Section 2.8.11 *The Best Matches Window* on page 42.

## 2.6.5.E The Components of the Document Tab

The **Document** tab components enable you to test one document using several operations. Use the information in the following table to determine how to use each component of this window.

Table 2-3: Document Tab Components

Field or button	Description
Test File	Specify one of the following operations: <ul style="list-style-type: none"> <li>- a path to a document</li> <li>- a URL to test a Web page</li> </ul>
Go	Begin loading the document.
Stop	Stop loading the document.
	Use the Open window that appears to locate the document on your machine that you want to test.
	When active, SAS Concept Creation is loading a Web page into the <b>Document</b> tab.
test file window	Use the test file window to perform one of the following operations: <p>See a tested document</p> <p>Double-click on a single document in the <b>Testing</b> tab and the <b>Document</b> tab appears. The matching definition terms are highlighted in red in the test document.</p> <p>Test a single document</p> <p>Access a text in the <b>Document</b> tab when you specify the path to the document using the <b>Test File</b> field or enter a URL.</p>
status window	See the status of the document, or the number of matches, for the selected concept. The status window is located to the left of the <b>TEST</b> button.
	Navigate through the matched concept terms in the tested document when you click the forward and backward buttons.
TEST	Test the loaded document.

Table 2-3: Document Tab Components (Continued)

Field or button	Description
Selected concept	Test only against the selected concept.
All concepts	Test this document against all of the concepts in this project. <b>Note:</b> Select this radio button and click <b>Test</b> . The Best Matches window appears. This window displays a list of concept matches ordered from best (top of the list) to worst (bottom).
Browser View 	Select this operation and the following buttons appear in the lower right-hand side of this interface: <ul style="list-style-type: none"> <li>- <b>Home:</b> Go to the home page.</li> <li>- <b>Back:</b> Return to the last viewed page.</li> <li>- <b>Forward:</b> Go to the next page.</li> <li>- <b>Refresh:</b> Update the Web page.</li> <li>- <b>Stop:</b> End the current process.</li> </ul>
Remove Tags	See the text of the selected Web page without any markup tags.

## 2.7 The Project Settings Windows

### 2.7.1 Project Settings Overview

Use the Project Settings windows to set taxonomy-wide operations. If you choose to develop a SAS Concept Creation project that uses more than one language, set the project settings for each language taxonomy separately. You can specify some of the project settings before you add concepts to the taxonomy. For more information, see Section 3.6 *Choosing Project Settings* on page 59. Modify these settings after testing or during the various stages of project development. For example, change your project settings if you do not obtain the testing results that you require.

---

To access Project Settings, complete this step:

Select **Project --> Settings**. The **LITI**, **Concordance**, and **Misc** tabs appear in the Project Settings interface.

*Display 2-8 Project Settings Tabs*



## 2.7.2 The LITI Tab

Use the selections in the **LITI** tab to specify how matching strings are treated.

### **Overlapping Concept Matches**

determine how SAS Concept Creation treats overlapping matches. Overlapping matches are strings where part, or all, of a matching string meets the match requirements for more than one concept.

#### **All matches**

return all of the terms that match any of the concept definitions in this project

#### **Longest**

return a match on the concept that matches the longest string.

#### **Best**

---

return the match with the highest priority setting, only.

---

**Note:** If all of the tested concepts have the same priority setting, only the longest matches are returned. For more information, see Section 4.4.17 *The Priorities and Project Settings* on page 75.

---

#### **Return all identical matches**

if you select either the **Longest** or **Best Matches, Return all identical matches** becomes available. Select this check box and SAS Concept Creation returns all of the identical longest or best matches—depending on your selection.

#### **Remove duplicate facts**

if you specify either a PREDICATE or a SEQUENCE rule, choose this operation to return only the first instance of a match. For more information, see Section 4.7.2 *The Predicate Sequence Example* on page 107 and Section 4.7.3 *The Predicate Examples* on page 109.

---

**Note:** These settings do not affect the returns specified by the REMOVE\_ITEM rule that excludes matches on a concept for disambiguation purposes. For more information, see Section 4.6.6 *Disambiguating Concepts* on page 89.

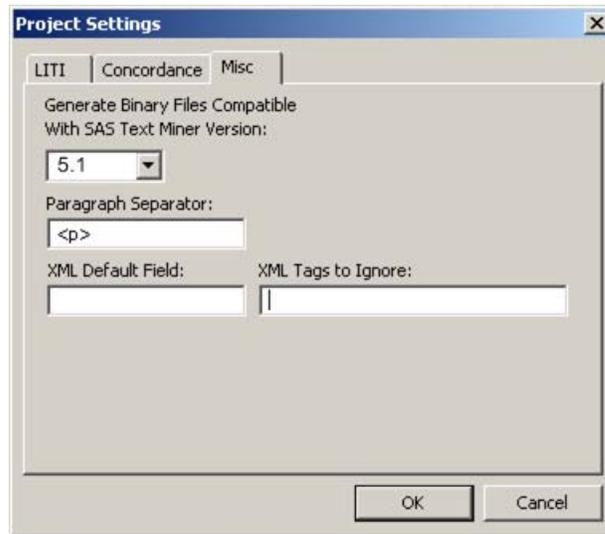
---

### 2.7.3 The Misc Tab

Use the **Misc** tab to specify the various settings that apply to the concepts extractor.

---

Display 2-9 Misc Tab



Use the **Misc** tab to specify the following settings:

**Generate Binary Files Compatible With SAS Text Miner Version**

if you are are not using SAS Text Miner 5.1, click  and select either 4.2 or 4.2M1.

**Paragraph Separator**

input the string that is used as a paragraph separator within your documents when you choose to use the `PARA` operator. For example, type `<P>`.

**XML Default Field**

limit search to the specified field. If you leave this field blank, all of the XML fields in the input XML document are searched.

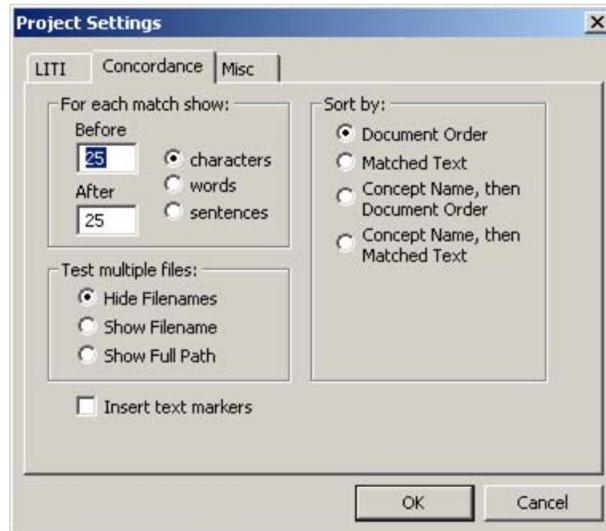
**XML Tags to Ignore**

choose to exclude one or more XML fields when processing your XML documents.

---

## 2.7.4 The Concordance Tab

Select the Project Settings - Concordance window to choose the display parameters for concept matches. (To see the concordance window, click **Concordance** in the **Document** tab.) A concordance provides a list of the matching terms in the document.



### For each match show

specify how many matching characters, words, or sentences are displayed in the concordance window:

#### **Before** (default: 25)

specify how many characters, words, or sentences to display before the match.

#### **After**

(default: 25) choose how many characters, words, or sentences to display after the match.

#### **characters**

(default) apply the numbers set in the **Before** and **After** fields to the letters in the alphabet, numbers, hyphens, and so on.

#### **words**

---

apply the numbers set in the **Before** and **After** fields to individual words.

**sentences**

return the specified number of sentences, set in the **Before** and **After** fields.

**Sort by**

classify matching terms in the concordance view of the **Document** tab:

**Document Order**

display the matches in the order in which the concepts occur in the document.

**Matched Text**

sort the matches alphabetically.

**Concept Name, then Document Order**

sort by matched concept name. Then sort by the order of appearances in the text.

**Concept Name, then Matched Text**

sort the matches by matched concept name and then sort these matches alphabetically.

**Test multiple files**

specify these operations when you use more than one testing file:

**Hide Filenames**

(default) do not show the names of the files that match in the concordance view.

**Show Filename**

display the test results, and to the right of this, the name of the file.

**Show Full Path**

display the test results with the name of the file. The full path of the file appears to the right of the results.

**Insert text markers**

display text markers in the concordance view of the **Document** tab when you test a single file against multiple concepts. The match text fields display the concept that is the best match for the matched term that is returned. An example of these tags is `<concept1>...</concept1>`.

---

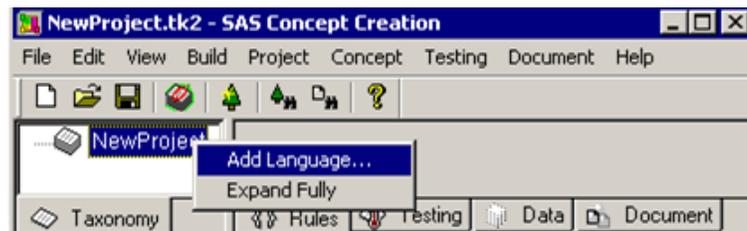
## 2.8 The Miscellaneous Windows

### 2.8.1 The Select a Language Window

Use the Select a Language window to choose the language for the entire taxonomy or a branch of your project. You can choose from any of the languages that you also install for SAS Text Miner.

To access and use the Select a Language window, complete these steps:

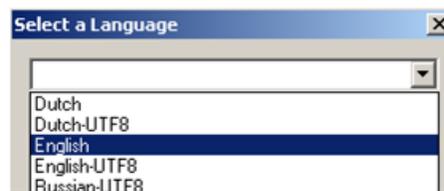
1. Right-click on the project node in the **Taxonomy** tab and select **Add Language** from the drop-down menu that appears.



The Select a Language window appears.



2. Click  to the right of the blank field and select a language that you purchased.



---

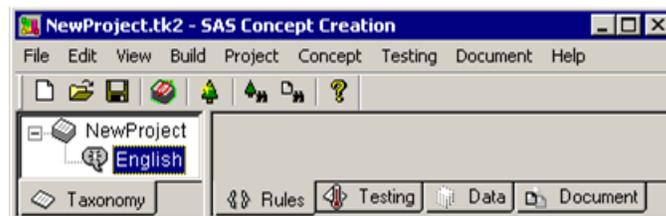
Languages followed by -UTF8 are in UTF-8 encoding. These languages include English, Chinese, Japanese, Korean, and Russian (Cyrillic characters), and so on. If the language is not followed by -UTF8, Latin-1 is used as the character set encoding.

---

**Notes:** When UTF-8 encoding is specified, test only documents that are UTF-8 encoded.  
If you use UTF-8 encoding, make sure that your computer has the appropriate language fonts installed.

---

3. Click **OK**. The selected language is added to your project. See the English example below.



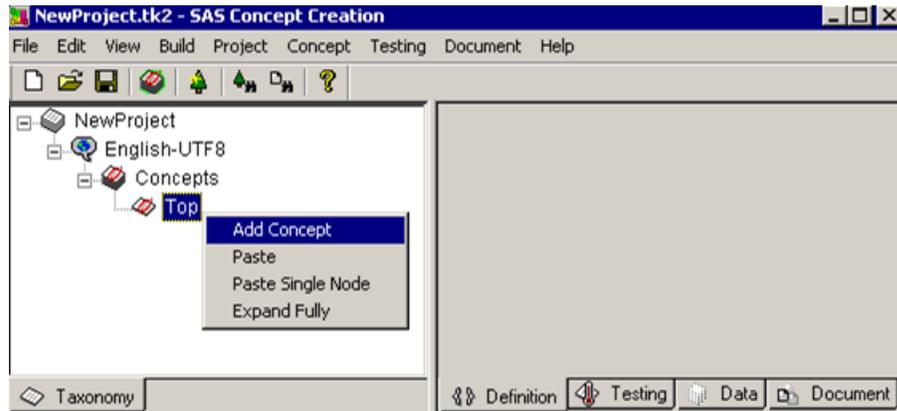
---

## 2.8.2 The Enter Names Window for UTF-8 Languages

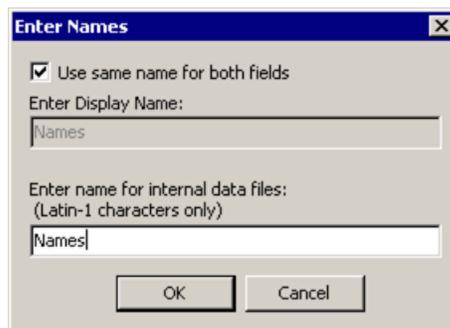
Concepts that use UTF-8 encoding require two names in the Enter Names window. Both of these names appear in the **Taxonomy** tab.

To access and use the Enter Names window, complete these steps:

1. Right-click on the **Top** node in the **Taxonomy** tab and select **Add Concept** from the menu that appears.



The Enter Names window appears.



2. (Optional) Select **Use same name for both fields** and the **Enter Display Name** field is dimmed and unavailable.

3. (Optional) Enter the name of the concept into the **Enter Display Name** field using UTF8 language characters.
4. Enter the name for your concept into the **Enter name for internal data files (Latin-1 characters only)** field.
5. Click **OK**. The new concept appears in the **Taxonomy** tab.

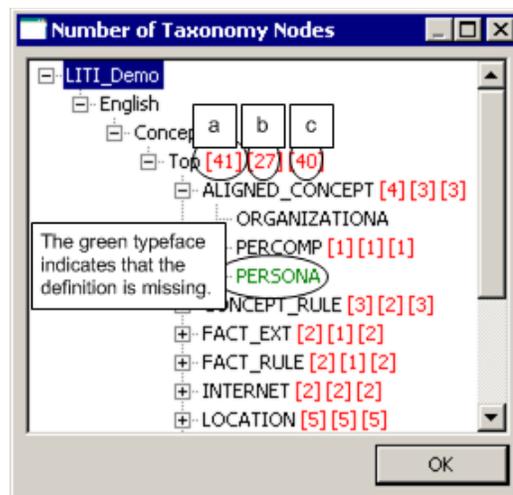
### 2.8.3 The Number of Taxonomy Nodes Window

See the following information about the taxonomy nodes in the Number of Taxonomy Nodes window:

- number of nodes
- number of subnodes
- nodes without a definition

To access and use the Number of Taxonomy Nodes window, complete the following steps:

1. Select **View --> Number of Taxonomy Nodes**. The Number of Taxonomy Nodes window appears.



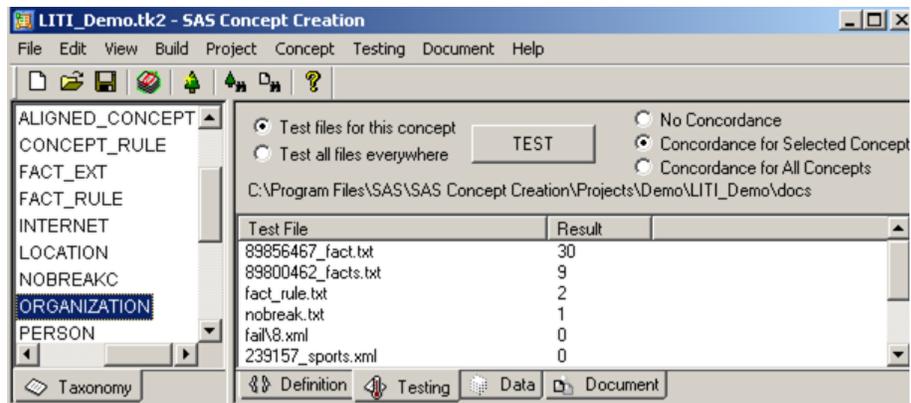
2. Use the Number of Taxonomy Nodes window to obtain the following types of counts (the list below correlates to the numbers in the figure above):
  - a. The number of taxonomy nodes represents all of the subnodes for the selected node in the **Taxonomy** tab. In the example above, 41 appears to the right of the `TOP` node.
  - b. The count of the children of the selected node that do not have subnodes is the second number that is displayed. In the example above, there are 27 children.
  - c. The number of subnodes that have a definition is the last count that is displayed. In the example above, `PERSONA` is highlighted in green because this node does not have a definition
3. Click **OK** to close this window.

## 2.8.4 The Concordance Windows That Are Available through the Testing Tab

There are two concordance windows that appear when you select a test operation with either of the available concordance operations.

To see the concordance matches in both windows, complete these steps:

1. After you load the testing documents for your project, click the **Testing** tab.



2. Select **Concordance for Selected Concept**.
3. Click **TEST**. The Concordance window appears.



4. See the terms that matched the selected concept inside the <Match>...</Match> tags.
5. Click **X** to close the Concordance window.
6. Select **Concordance for All Concepts**.
7. Click **TEST**. The Concordance window appears.



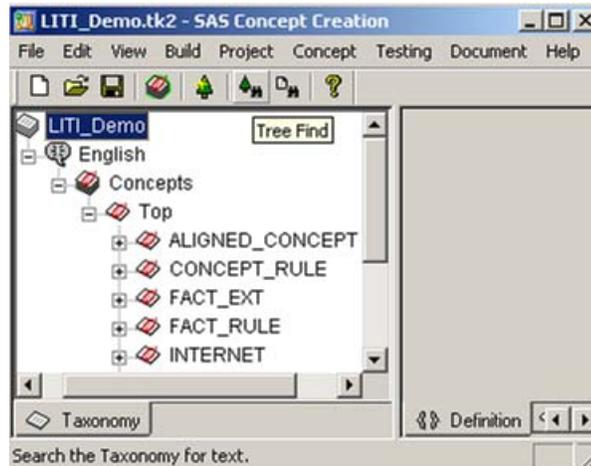
8. See the terms that matched the selected concept inside the tags for each concept. For example, see <COUNTRY>...</COUNTRY> that indicates a match on the COUNTRY concept.
9. Click **X** to close the Concordance window.

## 2.8.5 The Tree Find Window

Use the Tree Find window to locate a concept in a large taxonomy.

To find a concept, complete these steps:

1. Select the **Tree Find** icon on the standard toolbar.



The Tree Find window appears.



2. Enter the name of the concept that you want to locate into the **Find** field. For example, enter `FACT`.
3. (Optional) Select the **Match case** box to locate a matching term in the specified case. For example, type `FACT`.

---

**Note:** You can specify duplicate concept names, but only if their case is different.

---

4. Select **Find Next** to locate a match.

- 
- (Optional) Select **Replace** to access the Tree Replace window. For more information, see Section 2.8.6 *The Tree Replace Window* below.
  - Click **Cancel** to close this window.

## 2.8.6 The Tree Replace Window

Use the Tree Replace window to substitute a new name for the name that appears on one or more nodes in the **Taxonomy** tab.

To perform the replace operation, complete these steps:

- Select **Edit --> Tree Replace** and the Tree Replace window appears.



- Enter the text that you want to locate into the **Find** field.
- Enter the text that you want to substitute for the located term into the **Replace With** field.
- If you want to replace all of the original terms with the specified text, click **Replace All**.

---

**Note:** Use the **Replace All** button with care. This operation cannot be undone.

---

- Click **Cancel** to close this window. For more information, see Section 2.8.5 *The Tree Find Window* on page 36.

---

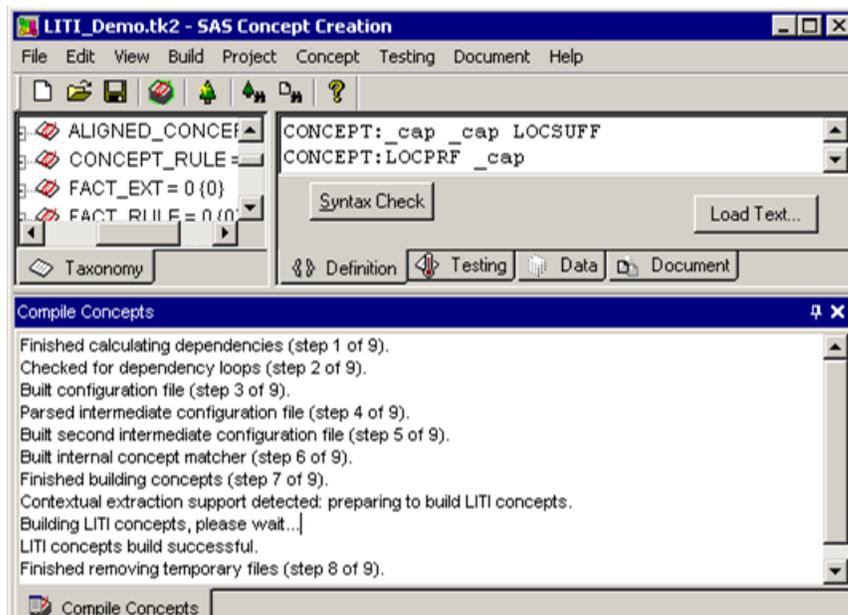
## 2.8.7 The Text Find and Replace Windows

Use the Text Find and the Text Replace windows like you use the Tree Find and Replace windows, or these operations in other applications. In SAS Concept Creation, these operations work in the **Definition**, **Testing**, and **Document** tabs.

## 2.8.8 The Compile Concepts Tab

The **Compile Concepts** tab appears at the bottom of the interface when you select **Build --> Compile Concepts**. This tab provides status information about the build process.

*Display 2-10 Compile Concepts Tab*



---

## 2.8.9 The Concordance Windows

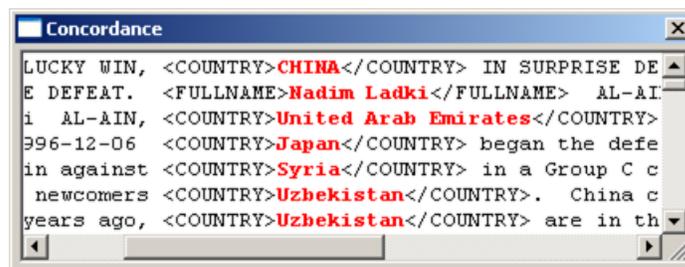
When you select either **Concordance for Selected Concepts** or **Concordance for All Concepts**, the Concordance window appears displaying the selected matches.

See the following examples:

*Display 2-11 Concordance for Selected Concept*



*Display 2-12 Concordance for All Concepts*

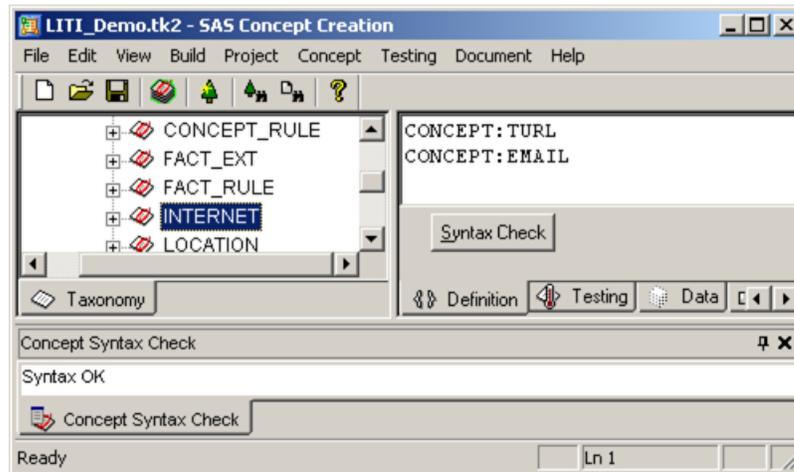


---

## 2.8.10 The Syntax Check Window

Click **Syntax Check** and the **Concepts Syntax Check** tab appears at the bottom of the user interface. This tab displays the results of the grammar check for the selected definition.

*Display 2-13 Syntax Check Tab*



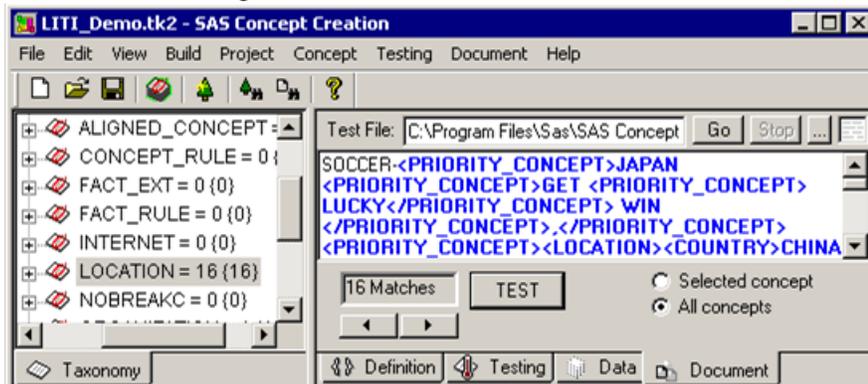
---

## 2.8.11 The Best Matches Window

Use the Best Matches window to see a list of the highest ranking concepts for your document. This window automatically appears when you select **All concepts** and click **TEST** in the Document window.

To access the Best Matches window, complete these steps:

1. Access a testing document in the **Document** tab.



2. Select **All concepts**.
3. Click **TEST**. The Best Matches window appears. See the example provided below.

The screenshot shows the "Best Matches" window, which is a small dialog box with a title bar and a close button. It contains a table with two columns: "Concept" and "Matches". The table lists the following concepts and their corresponding match counts:

Concept	Matches
Top/PRIORITY_CONCEPT	31
Top/LOCATION	16
Top/LOCATION/COUNTRY	16
Top/PERSON	8
Top/PERSON/LAST	4
Top/PERSON/NATIONMOD	4
Top/PERSON/FULLNAME	2
Top/PERSON/FIRST	2
Top/PERSON/TITLE	1
Top/ORGANIZATION	1
Top/ORGANIZATION/ORGCOMPND	1
Top/PERSON_EXPORT	1
Top/PERSON_EXPORT/SLN	1
Top/PERSON_EXPORT/SFN	1

4. Click **X** to close this window.

---

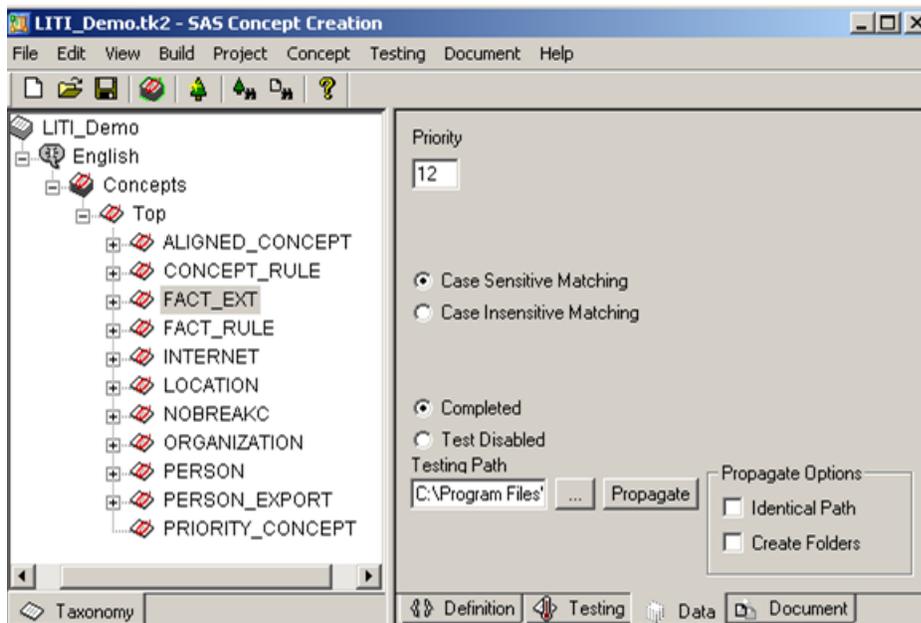
## 2.8.12 The Concept Priorities Window

The Concepts Priorities window displays the priority settings for concepts. Priority determines the matching concept when one input document matches two or more concepts and no other determiner makes one concept a better match than another.

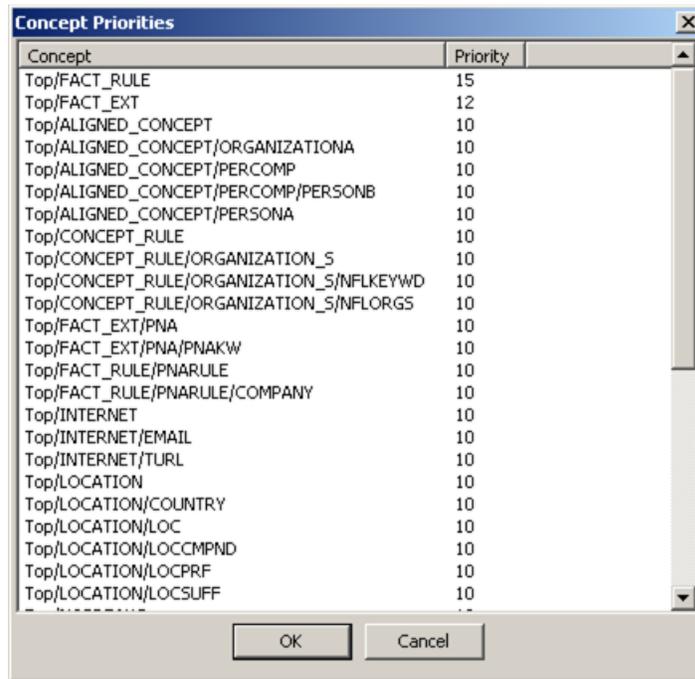
By default, this setting is set to 10. Increase this specification to make one concept rank higher than another when both are matched.

To access the Concept Priorities window, use these steps:

1. Specify a priority setting in the Data window for each concept that you want to rank. For example, type 12 into the **Priority** field for one concept and 15 into the **Priority** field for another concept.



- 
2. Select **Concept --> Priorities**. The Concept Priorities window appears.



The screenshot shows a dialog box titled "Concept Priorities" with a close button (X) in the top right corner. The dialog contains a table with two columns: "Concept" and "Priority". The table lists various concepts from a taxonomy, each with a numerical priority value. The concepts are listed in descending order of priority, with "Top/FACT\_RULE" having the highest priority of 15. Most other concepts have a priority of 10. At the bottom of the dialog, there are two buttons: "OK" and "Cancel".

Concept	Priority
Top/FACT_RULE	15
Top/FACT_EXT	12
Top/ALIGNED_CONCEPT	10
Top/ALIGNED_CONCEPT/ORGANIZATIONA	10
Top/ALIGNED_CONCEPT/PERCOMP	10
Top/ALIGNED_CONCEPT/PERCOMP/PERSONB	10
Top/ALIGNED_CONCEPT/PERSONA	10
Top/CONCEPT_RULE	10
Top/CONCEPT_RULE/ORGANIZATION_S	10
Top/CONCEPT_RULE/ORGANIZATION_S/NFLKEYWD	10
Top/CONCEPT_RULE/ORGANIZATION_S/NFLORGS	10
Top/FACT_EXT/PNA	10
Top/FACT_EXT/PNA/PNAKW	10
Top/FACT_RULE/PNARULE	10
Top/FACT_RULE/PNARULE/COMPANY	10
Top/INTERNET	10
Top/INTERNET/EMAIL	10
Top/INTERNET/TURL	10
Top/LOCATION	10
Top/LOCATION/COUNTRY	10
Top/LOCATION/LOC	10
Top/LOCATION/LOCCMPND	10
Top/LOCATION/LOCPRF	10
Top/LOCATION/LOCSUFF	10

3. See a ranked list of concepts according to the specified priorities. (By default, each **Priority** specification is set to 10.)
4. Click **OK** to close this window.

---

**Hint:** The Concept Priorities window does not specify matches. This window displays only the priorities for each of the concepts in the taxonomy.

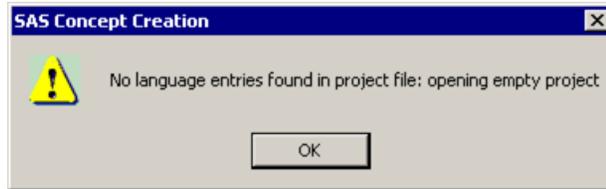
---

---

## 2.8.13 SAS Concept Creation Status Window Example

If you name and save a new project before you add a language, a SAS Concept Creation status window appears when you access this project.

*Display 2-14 SAS Concept Creation Status Window*



Click **OK** to close this window. Add a language to your project.

## 2.9 The Drop-down Taxonomy Node Operations

### 2.9.1 The Project Name Node Operations

Right-click on the first node that appears after you name your project. This is the name of the project.

*Display 2-15 Add Language and Expand Fully Operations*



#### **Add Language**

specify a language for this branch of your taxonomy. The Select a Language window appears with a drop-down list of the languages that you purchased. For more information, see Section 2.8.1 *The Select a Language Window* on page 31.

---

**Expand Fully**

see all of the nodes in this taxonomy.

## 2.9.2 The Language Node Operations

Right-click on the language node in your taxonomy in order to access the drop-down operations.

*Display 2-16 Language Node Drop-down Operations*



Use the following operations to change your taxonomy structure:

**Delete Language**

remove the language node for this taxonomy.

**Enable Concepts**

add the `Concepts` node and beneath it, add additional nodes.

**Expand Fully**

access the full taxonomy of nodes.

---

## 2.9.3 The Concepts Node Operations

Right-click on the `Concepts` node in order to access some of the taxonomy operations.

*Display 2-17 Concepts Node Operations*



Select from the following operations:

### **Remove Concepts**

delete this node from the taxonomy when you choose this operation.

---

**Warning:** When you select this operation, all of the child nodes below the language node are deleted with the `Concepts` node.

---

### **Expand Fully**

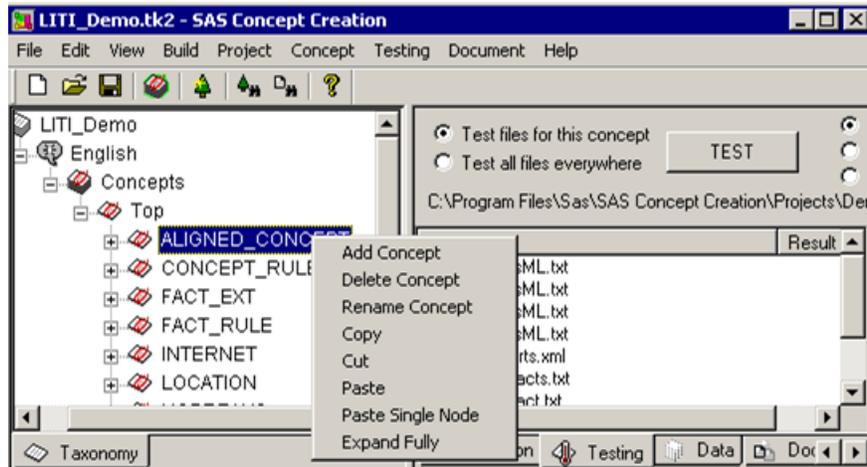
click to fully expand the selected branch of the taxonomy.

---

## 2.9.4 The Individual Concept Node Operations

Right-click on a concept node and a list of operations appears in the drop-down menu:

*Display 2-18 Individual Concept Operation*



Select from the following operations for concepts. The **Cut**, **Copy**, and **Paste** operations are self-explanatory:

**Add Concept**

add a child concept to the selected parent node.

**Delete Concept**

remove the selected concept.

**Rename Concept**

change the name of the concept.

**Paste Single Node**

paste one copied node as a child of the selected concept.

**Expand Fully**

access the selected branch of your taxonomy.

---

# Chapter: 3

## Creating Projects

---

- *Overview of Creating Projects*
- *Start SAS Concept Creation*
- *Create a New Project*
- *Saving the Project*
- *Access an Existing Project*
- *Choosing Project Settings*
- *Navigating through the Taxonomy*
- *Specify the .li File in SAS Text Miner*

### 3.1 Overview of Creating Projects

Build a `.li` file within the framework of a project. Each of the concepts in the project are displayed in a taxonomy structure. The taxonomy is the tree-like structure that alphabetically organizes the concept nodes.

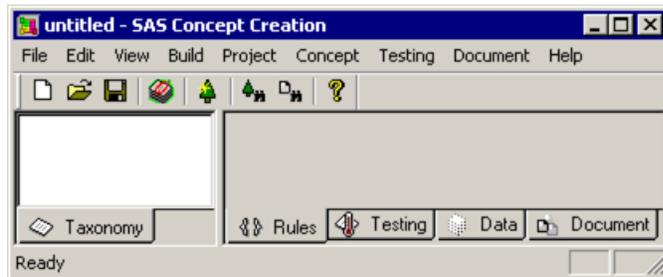
You write the definitions that define these concepts and test them using the **Testing** and **Document** tabs in order to ensure that these rules perform as expected. The concepts that you develop are built as a `.li` file that is used by the **Text Parsing** node of SAS Text Miner. SAS Text Miner applies these concepts as custom entities in real time.

---

## 3.2 Start SAS Concept Creation

To start SAS Concept Creation for SAS Text Miner (SAS Concept Creation), complete these steps:

1. Select **Start --> Programs --> SAS Concept Creation** and the untitled user interface appears.



2. See Section 3.3 *Create a New Project* below. To access an existing project, see Section 3.5 *Access an Existing Project* on page 57.

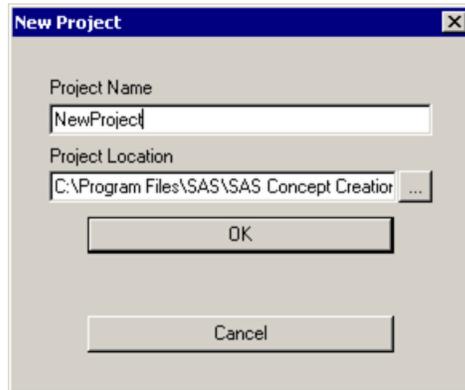
---

## 3.3 Create a New Project

Use this section to develop a new project the first time you use SAS Concept Creation. When you create a project, you define the concepts that are used as custom entities in SAS Text Miner.

To create a new project, complete these steps:

1. Select **File --> New Project** and the New Project window appears.

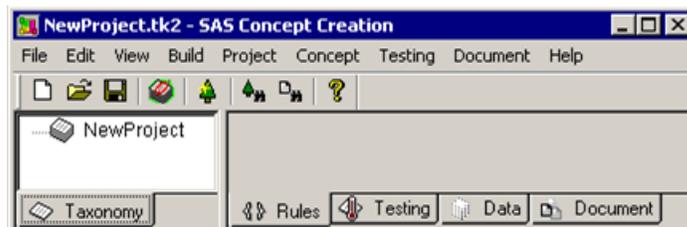


2. Enter the name of the new project into the **Project Name** field. For example, type `NewProject`.
3. (Optional) Click  to locate a directory and load this directory location into the **Project Location** field. The default location for a 32-bit machine running an English version of Windows is:  

```
c:\Program Files\SAS\SAS Concept Creation\Projects.
```

For a 64-bit machine, the default folder might be different. For example, the folder might be entitled `Program Files (x86)`.
4. Click **OK** to save this project to the selected location.

- 
- The newly named project node appears in the **Taxonomy** tab. For example, see the `NewProject` node below:

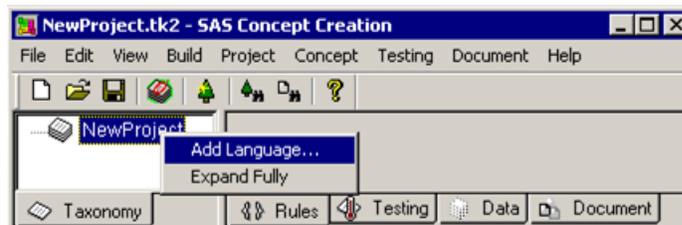


---

**Hints:** After you create a new project, set your project-wide settings. You can also choose to set your project-wide settings at a later stage in project development. For more information, see Section 3.6 *Choosing Project Settings* on page 59. Remember to save your project frequently throughout development. For more information, see Section 3.4 *Saving the Project* on page 56.

---

- Right-click on the project icon and select **Add Language** from the drop-down list that appears.



- 
7. The Select a Language window appears. Click  to the right of the blank field to select a language with an encoding for this language.



Languages that are represented in both Latin-1 and UTF-8, such as western European languages, have two entries in the drop-down list. All other languages use the multibyte character set encoding because UTF-8 can represent every character encoding in the Unicode character set.

---

**Notes:** If UTF-8 encoding is used, make sure that all of the testing and input documents are UTF-8 encoded. Also ensure that your computer has the proper language fonts installed.

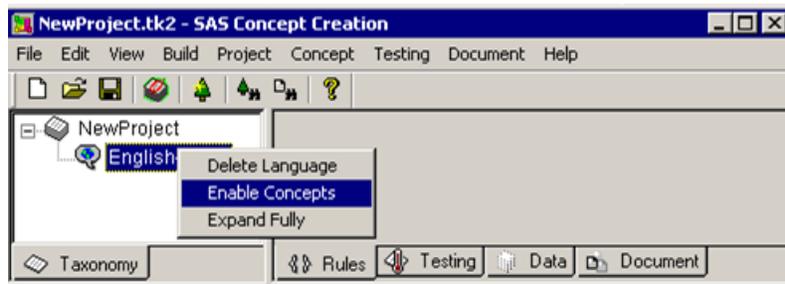
---

8. Click **OK**. The **Taxonomy** tab displays the new project node and the language node.

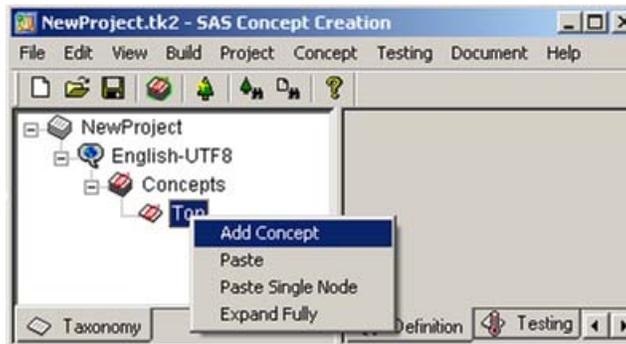


9. Right-click the language icon that appears in the **Taxonomy** tab. For example, right-click on `English-UTF8`.

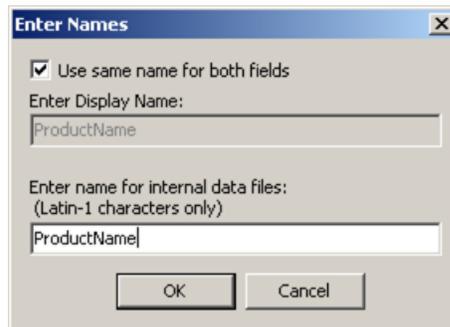
- 
10. Select **Enable Concepts** from the drop-down menu that appears.



11. Right-click on the **Top** node that appears, below the **Concepts** node, and select **Add Concept**.



- 
12. (For UTF-8 languages, only) The Enter Names window appears. See Section 2.8.2 *The Enter Names Window for UTF-8 Languages* on page 33.



If you add a concept in a language that is not UTF-8 enabled, the node is added to the taxonomy tree. See the window that can be used to enter the name of the node to the right of this node.

In either case, the node that appears looks similar to the example below. However, the second name is not included for nodes that appear for languages that are not UTF-8 encoded.



13. Select **File --> Save**. For more information, see Section 3.4 *Saving the Project* on page 56.
14. Continue adding nodes.
15. Write your concept definitions. For more information, see Section *Writing Concept Definitions* on page 65.

---

## 3.4 Saving the Project

### 3.4.1 Overview of the Save Operation

Use the Save operation to preserve the changes that you make to your project. By default, the project is saved before each test operation. However, it is important to manually save your project to preserve important changes or to create duplicate projects.

### 3.4.2 Manually Save an Existing Project

Manually save a project to keep different stages, or versions, of the project during development. The name of the project that appears in the title bar is the same name of the project folder that the application automatically creates. See the following example:

```
c:\Program Files\SAS\SAS Concept  
Creation\NewProject.tk2.
```

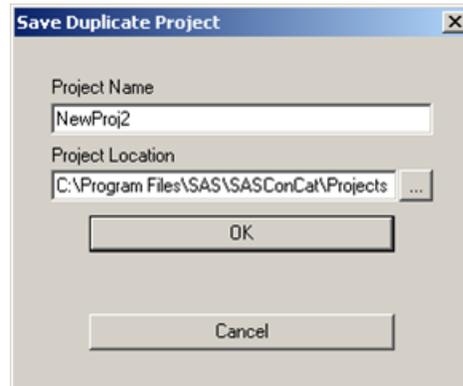
To save your project, select **File --> Save Project**.

### 3.4.3 Save a Duplicate Project

You can save your project as a duplicate project using another name. Use this operation when you want to preserve specific stages or versions of the project.

To create a duplicate project, complete these steps

- 
1. Select **File --> Save Project As**. The Save Duplicate Project window appears.



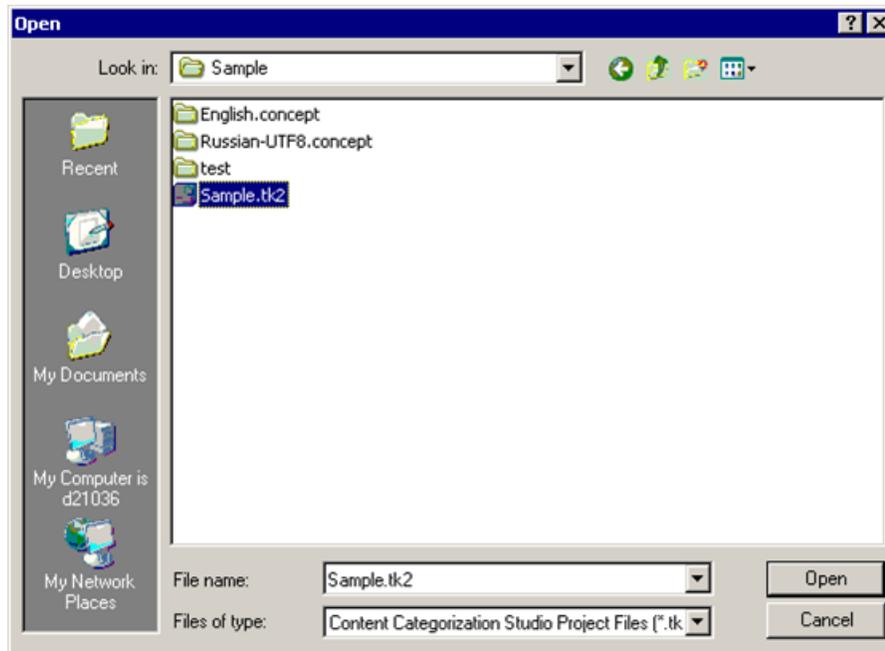
2. Enter the name of the duplicate project into the **Project Name** field. For example, type `NewProj2`.
3. (Optional) Click  to the right of the **Project Location** field to access the Select Directory window. Alternatively, leave the default project name and path that is automatically entered for you.
4. Click **OK**. The renamed project appears in the **Taxonomy** tab.

## 3.5 Access an Existing Project

After you create a project in SAS Concept Creation, you can access this project for further development or for reference purposes.

To access an existing project, complete these steps:

1. Select **File --> Open Project** and the Open window appears.



2. Click  to navigate through the program files and the `Projects` folder on your hard drive until you locate the `Sample.tk2` file.

---

**Hint:** The files for your projects are saved in a Windows folder that has the project name. For example, the files for the `NewProject` project are stored in the following directory:

```
c:\Program Files\SAS\  
    SAS Concept Creation\NewProject.tk2.
```

---

- 
3. Double-click the Sample project to access this project.



## 3.6 Choosing Project Settings

### 3.6.1 Overview of Project Settings

Use the Project Settings windows to specify matching, concordance, and other operations. These specifications apply only to the selected project. Many of these operations, such as those specified in the **LITI** tab affect how matching is applied to input documents.

### 3.6.2 Choose How Matches Are Returned

Use the **LITI** tab to specify how matches are returned when there are overlapping, identical, or duplicate matches.

To use the LITI window, complete these steps:

- 
1. Select **Project --> Settings** and the Project Settings dialog box appears.



2. Choose one selection under the **Overlapping Concept Matches** heading that determines how SAS Concept Creation treats overlapping matches. Overlapping matches are strings where part, or all, of the string matches more than one concept.
  - a. Leave the default selection **All matches** selected and SAS Concept Creation returns all of the terms that match any of the definitions in this project
  - b. Select **Longest** to return the longest match for the definition.
  - c. Select **Best** to return only the match with the highest priority setting.

---

**Note:** If all of the tested concepts have the same priority setting, only the longest matches are returned.

---

3. If you select either the **Longest** or **Best Matches**, **Return all identical matches** is available. Select this check box and SAS Concept Creation returns all of the identical longest or best matches.

If you specify either a `PREDICATE` or a `SEQUENCE` rule, choose this operation to return only the first instance of a match. For more information, see

---

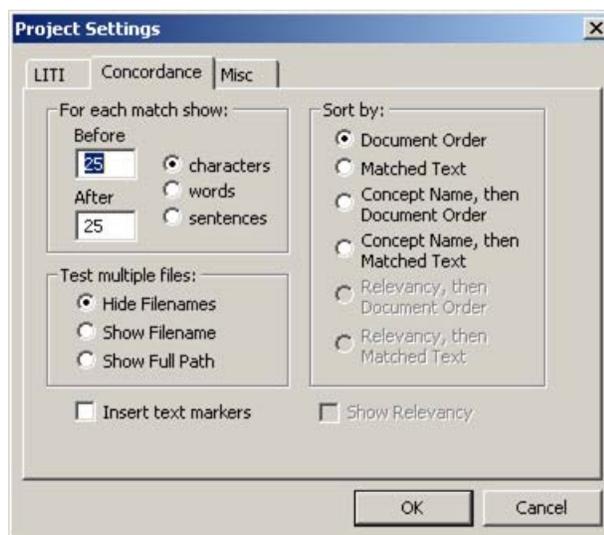
Section 4.7.2 *The Predicate Sequence Example* on page 107 and Section 4.7.3 *The Predicate Examples* on page 109.

4. Click **OK**.

### 3.6.3 Choose the Concordance Operations

Set the project-wide settings for the concordance operation. This operation displays the matched terms in input documents according to the specifications that you set here. Specify these settings in the **Concordance** tab.

*Display 3-1 Concordance Default Settings*



To specify settings in the **Concordance** tab, complete these steps:

1. Set all of the settings that are relevant at this time. For more information, see Section 2.7.4 *The Concordance Tab* on page 29.
2. Click **OK**.
3. Select **Build --> Compile Concepts**.
4. Select **File --> Save**.
5. Begin testing the concepts.

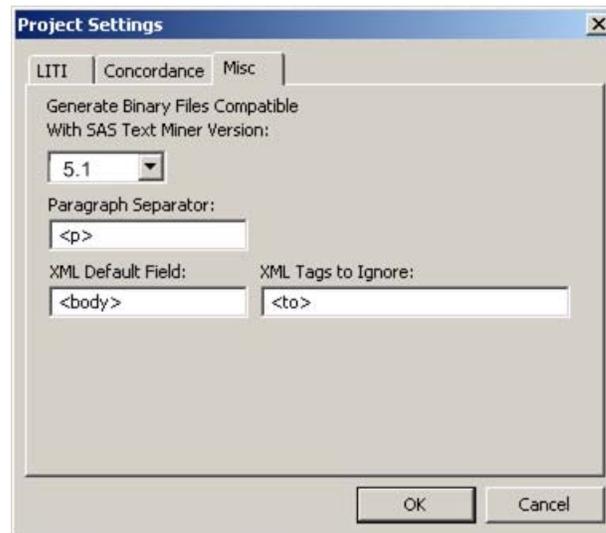
---

### 3.6.4 Choose Miscellaneous Operations

The **Misc** tab contains project-wide settings that affect the application and enable you to specify the paragraph separators that are found in input texts.

To specify settings in the **Misc** tab, complete these steps:

1. Click the **Misc** tab.

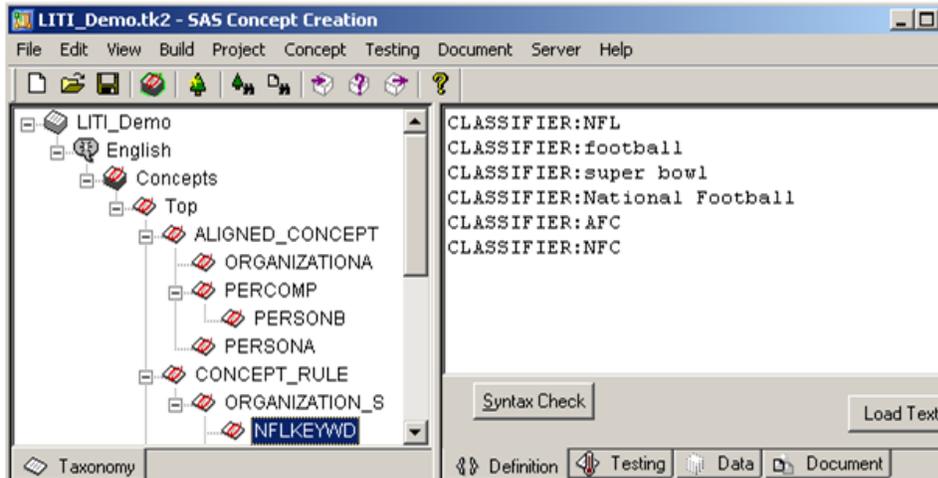


2. If you are not using SAS Text Miner 5.1, click  and select either 4.2 or 4.2M1.
3. Enter the paragraph separators used in your input documents into the **Paragraph Separator** field. For example, type <p>.
4. Enter the default field to search in XML documents into **XML Default Field**. For example, type <body>.
5. Enter the tags to overlook when searching XML documents into **XML Tags to Ignore**. For example, type <to>.
6. Click **OK**.

## 3.7 Navigating through the Taxonomy

After you create concepts, the **Taxonomy** tab displays a hierarchical view of the individual concepts that comprise your taxonomy. You can use standard Windows controls to navigate through, and to manipulate, these taxonomy nodes. See the following definition example:

*Display 3-2 Concepts Displayed in the Taxonomy Tab*



`Top` is the permanent name for the first node in the concept hierarchy in the **Taxonomy** tab. Every concept below `Top`, such as `ALIGNED_CONCEPT` or `PERSONA`, is a child of the `Top` node. These concepts, in turn, can also be the parents of other subcategories or children. For example, `ORGANIZATION_S` is the parent of the child categories `NFL_KEYWORD` and the child of `CONCEPT_RULE`.

The `PERCOMP` concept has a  next to it. This sign indicates that `PERCOMP` has one or more subcategories that are now displayed.

---

## 3.8 Specify the .li File in SAS Text Miner

After you create a project, SAS Concept Creation outputs a `.li` file. To use the `.li` file that SAS Concept Creation generated, set these properties in the **Text Parsing** node of SAS Text Miner:

1. Set the **Find Entities** property to **All** or **Custom**.
2. Enter the path to the `.li` file in the **Custom Entities** field.

The entities that are created in SAS Concept Creation are defined as *custom entities* in SAS Text Miner.

---

# 4

## Writing Concept Definitions

---

- *Overview of Definitions*
- *Before You Write Your Definitions*
- *The Rule Types*
- *The Building Blocks*
- *The Operators*
- *Some Rule Examples*
- *Locating Facts*
- *The Coreference Operators*
- *XML Fields in Rules*
- *Writing Multiple Rules for One Definition*
- *Troubleshooting Your Rules*

### 4.1 Overview of Definitions

Use SAS Concept Creation (SAS Concept Creation) to develop the concepts that are used as custom entities in SAS Text Miner:

- Write a simple rule that matches one term specified in a list of entries.
- Locate a match for a unique concept where each individually specified term in the concept appears in one of the rules that together define this concept.
- Match a concept if it appears within the specified context, only.
- Locate multiple partial matches and return them as full concept matches. These matches can occur only if there is a match on the fully defined concept within the input document.

- 
- Write restrictive rules to prevent matches from occurring within specified contexts.
  - Disambiguate matches. Avoid possible matches on concepts that are specified using identical terms with different meanings.
  - Specify part-of-speech tags to locate concepts.
  - Use Boolean operators and various types of operators to increase the matching precision of your rule.
  - Specify case-sensitive rule matches in the Data window.
  - Use the stemming operator to return all of the forms of a word. Alternatively, choose to return only the noun or verb forms of the word.
  - Specify coreference operators for pronoun resolution. In other words, when a pronoun or another word refers to the canonical form for a term, return the canonical form.
  - Use the `PRIORITY` setting to specify that one rule is weighted more than another and to prevent the return of false positives for coreference matches. (In other words, you can rank one rule higher than another rule within the same definition.)
  - Match predicates by specifying multiple arguments to extract a fact.
  - Identify the semantic relations between concepts by using predicate rules with logical operators.
  - Specify XML fields to limit matches to these fields.
  - You can write comments into your rules.

---

## 4.2 Before You Write Your Definitions

Consider the following information before you write your concept definitions:

- *Concepts* is another word for the term *custom entities* that is used in SAS Text Miner. SAS Text Miner identifies standard entities such as Percent, Phone, Time, and so on. In order to identify custom entities, SAS Text Miner uses the concepts that you define in SAS Concept Creation.
- The terms *rule* and *definition* are used interchangeably. Properly speaking, definitions apply to all of the rules for one concept.
- Rule types, for example `CLASSIFIER` and `C_CONCEPT`, are written using uppercase letters.
- By default, SAS Concept Creation performs case-sensitive matching.
- By default, the **Priority** setting in the Data window is set to 10. You can also specify a `PRIORITY` setting that overrides this setting within some rules.
- By default, matches can occur in any part of an input document. When the `PARA` or various `SENT` operators are specified, a match is returned if the matches occur in one paragraph, sentence, or the specified number of sentences.
- The settings in the Project Settings - LITI dialog box can affect match returns.

---

## 4.3 The Rule Types

There are many types of definitions. You can also specify more than one rule for each of your definitions. A match on the concept occurs if there is a match on any one of these rules.

### CLASSIFIER

Specify lists of terms where each classifier rule consists of the word `CLASSIFIER` followed by a string. For more information, see Section 4.6.1 *The Classifier Rules* on page 83.

### CONCEPT

Reference one or more concepts and use the `_cap` term to specify that a match only occurs on a word that begins with an uppercase letter. When more than one concept is referenced, a relationship is specified between the matching terms. You can also use `CONCEPT` rules to locate, or to discover, related information. For more information, see Section 4.6.2 *The Sequence of Classifier Entries* on page 84.

### C\_CONCEPT

Specify the order for the match components in an input document using these definitions. For more information, see Section 4.6.3 *Context Matching* on page 85.

### NO\_BREAK

Prevent partial matches on a rule that is specified within this definition. Use this rule to determine that an entire phrase is treated as a single word. For more information, see Section 4.6.5 *Eliminating Partial Matches* on page 88.

### REMOVE\_ITEM

Eliminate a false match in input documents where one word is a unique identifier for two concepts. This rule ensures that the correct context for the match is considered. For more information, see Section 4.6.6 *Disambiguating Concepts* on page 89.

### REGEX

Match information that follows a preset pattern. For more information, see Section 4.6.10 *The Regular Expressions in a Definition* on page 97 and Appendix A .

### CONCEPT\_RULE

---

Specify Boolean operators to increase precision (relevancy of the matches) and recall (return all matching texts). For more information, see Section 4.6.11 *The Sentence Operator in a Definition* on page 98.

SEQUENCE

Extract facts from input documents if the facts appear in the order specified. For more information, see Section 4.7.2 *The Predicate Sequence Example* on page 107.

PREDICATE\_RULE

Specify the arguments that define your facts. Facts are related pieces of information in a text that are often located and matched as phrases. For more information, see Section 4.7.3 *The Predicate Examples* on page 109.

## 4.4 The Building Blocks

### 4.4.1 Overview of the Building Blocks

SAS provides `n-gram` sequence features that are often used in natural Language Processing (NLP). These sequences specify the context that is necessary for the specified concept to match. Before you write your rules, consider the building blocks that are explained in this section.

### 4.4.2 Case-Insensitive Matching

By default, SAS Concept Creation applies rules to input documents in a case-sensitive manner. You can specify case-insensitive matching when you click **Case Insensitive Matching** in the **Data** tab. This setting applies to the entire definition of the selected concept, only.

### 4.4.3 Entering Comments into Rules

Any character, or characters, following the pound sign (#) are considered to be comments. For a literal # to match, it should be escaped as \#.

---

#### 4.4.4 The Tokens

Add tokens to your definitions:

- words, including noise words such as *and*, *the*, and *a*
- numbers including date and time
- newline mark
- URLs

Specify an undetermined token using the `_w` term. When you specify this term, SAS Concept Creation returns a match on any word that occurs in this position in the document. If, on the other hand, there is an exact token that you want this concept to match, you can specify this word in any concept rule. When tokens are specified in `CONCEPT_RULES` and `PREDICATE_RULES`, these tokens are set off with quotation marks (""). For more information, see Section 4.4.6 *The \_w Term* on page 70.

#### 4.4.5 The `_c` Marker

Use the context marker (`_c`) to specify that a match is returned if the keyword is located within the specified context. For example, you can match any `COMPANY` concept that is immediately followed by the term *New York*:

```
C_CONCEPT:_c{COMPANY} New York
```

You can also use this marker to locate and return known and unknown words. See the following two examples:

```
C_CONCEPT:COMPANY _c{New York}
C_CONCEPT:COMPANY _c{cap}
```

#### 4.4.6 The `_w` Term

Use the word term (`_w`) to specify that a match can occur on a word. For example, you can match any type of business. This is true if `_w` immediately follows a reference to the `COMPANYTYPE` concept:

```
C_CONCEPT:_c{COMPANYTYPE} _w
```

This example could also return a match on law *firm*.

---

**Hint:** The `_w` term matches any single term. A term can consist of alphabetic or non-alphabetic characters. For example, *today*, *<*, *Web*, *1.0*, and so on.

---

#### 4.4.7 The `_cap` Term

Use the `_cap` term in ways that are similar to the `_w` term. However, `_cap` only returns matches on words that begin with an uppercase letter. Use `_cap` to locate an unknown term that begins with an uppercase letter, or to match a single upper case letter. Alternatively, specify this term multiple times. When you repeatedly specify `_cap`, you can locate all of the unknown, consecutive occurrences of words that begin with an uppercase letter. This term can be used with all of the rule types except for the `CLASSIFIER` and `REGEX` rules. You can also replace `_w` with `_cap` in the example provided for Section 4.4.6 *The `_w` Term* above. In this case, the word *Firm*, or another word beginning with an uppercase letter, is a match.

#### 4.4.8 The `>` Symbol

Documents often reference a unique, full string only once. After that these references might be made by one word from the original string. Use the greater than (`>`) symbol with either the `C_CONCEPT`, `OR CONCEPT_RULE`, or a coreference operator (`_ref`). For more information about coreference, see Section 4.8.3 *How to Use the `_ref` Operator with the `>` Symbol* on page 115. Every occurrence of the bracketed term is a match if the entire rule is matched at least once in the input text.

Specify the greater than symbol within the `C_CONCEPT` rule using the `_c{ }>` syntax. For example, use this symbol to specify that every instance of the last name *Pelosi* is returned as a match after the entire term *Ms. Nancy Pelosi* is located. See the following example where `TITLE` and `FIRST` refer to classifier concepts with a list of titles and first names, respectively:

```
C_CONCEPT:TITLE FIRST _c{ _cap}>
```

---

### 4.4.9 The Quotation Marks

Use quotation marks (" ") to enclose tokens and concepts when writing a `CONCEPT_RULE`, `REMOVE_ITEM`, or `PREDICATE_RULE`. This example returns a match on *Mount Washington* if the term *Mount*, and a match on the concept `NAME`, appear within seven words of a match on the `STATE` concept:

```
CONCEPT_RULE: (DIST_7, "_c{Mount NAME}", "STATE")
```

### 4.4.10 The Parentheses, Square Braces, and Curly Braces

Use parentheses ( ( ) ), square braces ( [ ] ), and curly braces ( { } ) as appropriate. These symbols qualify the matches for all of the definitions except the `CLASSIFIER` and `CONCEPT` types.

Use parentheses ( ( ) ) to group the elements that comprise `CONCEPT_RULE`, `REMOVE_ITEM`, `SEQUENCE`, and `PREDICATE_RULE` definitions. For example, use parentheses with arguments and logical operators. Parentheses are also used with the `AND`, `OR`, `SENT`, `DIST_n`, `ORDDIST_n`, and `ALIGNED` Boolean operators. These operators are followed by a comma ( , ) and a space.

Use square braces ( [ ] ) to group `REGEX` rule elements with the `Export` operation. For more information, see Section 4.4.15 *The Export Feature* on page 74.

Use curly braces to delimit the information that is returned as a match. Curly braces ( { } ) are used with or without parentheses ( ( ) ), depending on the type of definition that you write. For more information, see Section 4.7.2 *The Predicate Sequence Example* on page 107 and the following example:

```
CONCEPT_RULE: (SENT, "_c{FIRST, _cap}",  
"TITLE", "COMPANY")
```

### 4.4.11 The Commas

Commas ( , ) always follow definition elements:

- Boolean operators are enclosed in parentheses ( ( ) ) and a space follows the comma ( , ) after this string.

- 
- Quotation marks (") enclose concept names and a comma follows the second quotation mark.
  - Separate the arguments used to construct facts with commas.
  - Commas follow logical operators in a PREDICATE\_RULE.

#### 4.4.12 The Colons

Use a colon (:) in the following cases:

- Type a colon after specifying the concept rule type. For example, use a colon with these rules CONCEPT, CLASSIFIER, and CONCEPT\_RULE.
- Use a colon when specifying terms to export to CLASSIFIER rules. For more information, see Section 4.6.7 *Exporting Classifiers* on page 91.
- Use colons between arguments for a SEQUENCE OF PREDICATE\_RULE concept. For more information, see Section 4.7.2 *The Predicate Sequence Example* on page 107 and Section 4.7.3 *The Predicate Examples* on page 109.
- Type a colon before a part-of-speech tag. For example, type :prep and :sep. For more information, see Section 4.6.9 *The Part-of-Speech Tags in a Definition* on page 96.

#### 4.4.13 The Spaces

When you write CONCEPT, CONCEPT\_RULE, or C\_CONCEPT definitions, type at least one space before each of the following items, tokens, concepts, part-of-speech tags, \_w terms, and \_cap terms. Also type a space before the \_c marker if it is preceded by a token, comma (,), or the name of a concept. See the following example:

```
CONCEPT_RULE:(ORDDIST_9, "_c{_cap} :sep _cap :sep and
                 _cap", "ORGTYP")
```

---

## 4.4.14 The Part-of-Speech Tags

Specify part-of-speech tags when you don't know the exact word that you are seeking. For example, `:Prep` to represent preposition and `:sep` to specify a separator character. A separator character is any punctuation mark. These part-of-speech tags are preceded by a colon (`:`) and a space. In addition, a space also precedes each of these tags. See the following example:

```
CONCEPT_RULE:(SENT, "_c{VACATION :Prep _cap :sep
LOCATION}", "vacation")
```

For a complete list of part-of-speech tags, see Appendix B.

---

**Note:** Use the part-of-speech tags that are listed in Appendix B. Do not use the part-of-speech tags that are used with SAS Text Miner. The part-of-speech tags that are specified in your definitions are mapped to those in SAS Text Miner at the time of application.

---

## 4.4.15 The Export Feature

Export a matched term to one or more concepts. Use the `Export=` operation to define a term that matches a classifier concept. Also use the coreference operator (`_ref`) with the export symbol to eliminate false positives. You can specify this operation within the concept definition. Alternatively, declare an acronym as part of the definition for the concept where the selected term is exported. See the following example:

```
FULLNAME: CLASSIFIER: [export=eLN:Clinton]: Bill Clinton
LASTNAME: eLN
```

The term `Clinton` is exported to the `LASTNAME` concept. When you write the export operation into a classifier rule, all instances of partial matches such as *Clinton* are returned. For this reason, the export feature functions in ways that are similar to the effects of placing the greater than symbol (`>`) at the end of a rule. For more information, see Section 4.6.7 *Exporting Classifiers* on page 91.

---

## 4.4.16 The Regular Expressions

Match known patterns by using regular expressions to specify a range of letters or numbers inside square braces ( [ ] ). For example, place `a-z` or `0-9` inside square braces. This specification matches any word beginning with an ASCII character whose value is between a and z, or numbers between 0 and 9 inclusive. You can also add a plus (+) sign after the last square brace. See the following example:

```
REGEX: [a-z]+
```

When you add the plus sign, all instances of terms beginning with a lowercase letter from the English alphabet match all occurrences of a word in the input document. You can continue to build this definition by specifying a context for the word occurrence.

You can also add either the % symbol or write out `percent`, after these bracketed numbers. This feature enables you to locate percentage matches in your documents. See the following example:

```
REGEX: [0-9]+%  
REGEX: [0-9]+ percent
```

This regular expression specifies that only numbers followed by the percentage sign match. In this example, either 99%, or 50 percent, are matches.

For more information, see Appendix A *Regex Syntax and Part-of-Speech*.

## 4.4.17 The Priorities and Project Settings

### 4.4.17.A Overview of Priorities

Priorities determine the concepts that are matched when priorities are applied to input documents by SAS Content Categorization Server. Matching is displayed in the Document pane and is applied after the concepts are uploaded to SAS Content Categorization Server as binary files.

For example, you might have a document that contains matches for both concept A and concept B. To prioritize a match on concept A, set the **Priority** setting in the **Data** tab for concept A to a higher number than that of concept B. Alternatively, you could specify `PRIORITY=n` in one or more rules in your definitions.

---

The `PRIORITY` rule specification that is set higher than 10, overrides the **Priority** setting in the **Data** tab. By default, the **Priority** setting in the **Data** tab is set to 10. For this reason, a `PRIORITY` setting in a rule ranks overlapping rule matches in one concept definition as well as matches on different concept definitions. For more information, see Section 4.6.8 *Setting Priorities for Overlapping Matches* on page 94 and Section 4.8.7 *Rank Coreference Definitions and Eliminate False Positives* on page 119.

See the following example where 35 overrides the default **Priority** setting of 10 in the Data window:

```
C_CONCEPT:PRIORITY=35: _c{CITY COUNTRY}
```

#### 4.4.17.B Choose Project Settings

Use the **LITI** tab in the Project Settings window to choose the types of matches that you want to return. These settings are particularly important when you specify priorities and when multiple matches occur within one input document.

To specify Project Settings, complete these steps:

1. Select **Project --> Settings** and the Project Settings window appears.



2. Select **All matches** to return matches on all of the matching rules in an input document.
3. Select **Longest** to return only the match with the most characters.
4. Select **Best** to return only the best match.

- 
5. Select **Return all identical matches** when you want to locate each instance of a rule match.
  6. If you specify either a `PREDICATE` or a `SEQUENCE` rule, you can select **Remove duplicate facts** to return the first instance of a match, only.

Priorities also affect concept matching. For more information and examples, see Section 4.6.14 *The ORDDIST Operator in a Definition* on page 104 and Section 4.8.7 *Rank Coreference Definitions and Eliminate False Positives* on page 119.

---

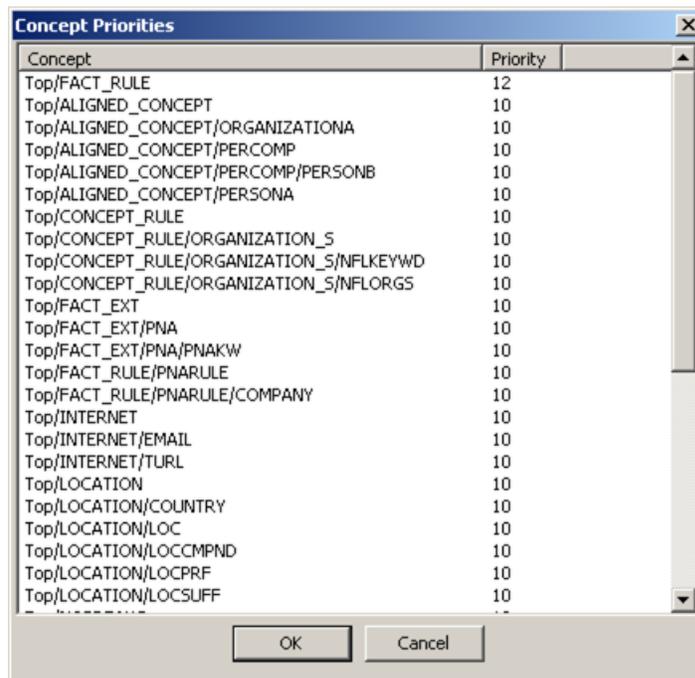
#### 4.4.17.C Seeing the Priorities for the Taxonomy

When you specify individual priority settings for one or more concepts in the Data window, you can see these results in the Concept Priorities window. Use this window to determine how ranking is applied internally by SAS Concept Creation when a document matches two or more concepts.

This window provides an overview of the priority settings so that you know what concepts are prioritized when an input document matches two or more concepts. The Concept Priorities window does not display the test results.

To use the Concept Priorities window, complete these steps:

1. Select **Concept --> Priorities**. The Concept Priorities window appears.



Concept	Priority
Top/FACT_RULE	12
Top/ALIGNED_CONCEPT	10
Top/ALIGNED_CONCEPT/ORGANIZATIONA	10
Top/ALIGNED_CONCEPT/PERCOMP	10
Top/ALIGNED_CONCEPT/PERCOMP/PERSONB	10
Top/ALIGNED_CONCEPT/PERSONA	10
Top/CONCEPT_RULE	10
Top/CONCEPT_RULE/ORGANIZATION_S	10
Top/CONCEPT_RULE/ORGANIZATION_S/NFLKEYWD	10
Top/CONCEPT_RULE/ORGANIZATION_S/NFLORGS	10
Top/FACT_EXT	10
Top/FACT_EXT/PNA	10
Top/FACT_EXT/PNA/PNAKW	10
Top/FACT_RULE/PNARULE	10
Top/FACT_RULE/PNARULE/COMPANY	10
Top/INTERNET	10
Top/INTERNET/EMAIL	10
Top/INTERNET/TURL	10
Top/LOCATION	10
Top/LOCATION/COUNTRY	10
Top/LOCATION/LOC	10
Top/LOCATION/LOCCMPND	10
Top/LOCATION/LOCPRF	10
Top/LOCATION/LOCSUFF	10
.....	..

2. See a ranked list of concepts according to the priorities that you specified in the Data window. (If you did not specify any priorities, this window does not display any concepts.)
3. Click **OK**.

---

## 4.5 The Operators

### 4.5.1 The Boolean Operators

To locate related information with greater precision, specify Boolean, or logical operators, with some types of rules.

Table 4-1: Boolean Operators

Operator	Description
<u>ALIGNED</u>	Disambiguate between matches on two concept rules. Disambiguation enables SAS Concept Creation to determine the correct match based on context. When terms are disambiguated, only one match is returned.
<u>AND</u>	Specify that a match can occur only when both arguments are present, somewhere within the entire document.
<u>OR</u>	Specify that if one, but not both, of the concepts or tokens is located a match is returned.
<u>DIST_n</u>	Specify the number of words between matches on rule terms. The first match takes the starting position 1, while the last match falls at or before the specified number of words.
<u>ORDDIST_n</u>	Specify the maximum word count between arguments. Otherwise this operator functions like the <code>DIST</code> operator above.
<u>SENT</u>	Specify a sentence delimiter. For example, <code>..</code> , <code>?</code> , or <code>!</code> . A match is returned when all of the specified components are located in the sentence where the first match occurs.
<u>SENT_n</u>	Specify a sentence delimiter that returns matches on multiple sentences.
<u>SENTSTART_n</u>	Specify that matches are returned within <code>n</code> words from the start of the sentence.
<u>SENTEND_n</u>	Specify that matches are returned within <code>n</code> words from the end of the sentence.

Specify a comma (,) and a space after a Boolean operator and enclose it in parentheses (()). For example, write `(SENT, "NAME")`.

---

#### 4.5.1.A The ALIGNED Operator

Use the `ALIGNED` operator to refer to a term that matches two concepts within one rule. The presence of this operator enables SAS Concept Creation to determine what concept is an exact match for this term.

For example, the following rule specifies that if a term matches both the `LOC` and `PERSON` concepts, only a match for the `PERSON` concept is returned. Matches for the `LOC` concept, such as `Washington`, are returned as a match on the `PERSON` concept:

```
REMOVE_ITEM:ALIGNED, ("_c{LOC}", "PERSON")
```

#### 4.5.1.B The AND Operator

Specify the `AND` operator for two or more arguments. A match only occurs if both arguments are present. For example, the following rule limits matches to `Bills` in documents where the word `football` also occurs:

```
CONCEPT_RULE:(AND, "_c({Bills})", "football")
```

#### 4.5.1.C The OR Operator

Specify the `OR` operator for two or more matched rule components. A match occurs for an input document if at least one of these components is present. For example, the following rule matches if either the token `Barack` or `Obama` is present in the text:

```
CONCEPT_RULE:(OR, "_c{Barack}", "_c{Obama}")
```

#### 4.5.1.D The DIST\_n Operator

Specify the maximum distance, in words, between located terms in order for a match to be returned for the selected concept. For example, if you want to specify that a match on the `FULLNAME` concept that appears within eight words of *Harvard University* is a match, write the following definition:

```
CONCEPT_RULE:(DIST_8, "_c{FULLNAME}",  
                "Harvard University")
```

---

#### 4.5.1.E The ORDDIST\_n Operator

Specify the order and distance between the terms or concepts that you want the selected concept to match. This operation locates and returns a match even when the usual contextual clues provided by adjacent matches are missing. For example, a match can be located when name and position do not follow one another. The following example returns a match on the POSITION concept when it is followed by the word `Obama`. This is true only if the term `Obama` is located within 12 words from a match on the POSITION concept.

```
CONCEPT_RULE: (ORDIST_12, "_c{POSITION}", "Obama")
```

#### 4.5.1.F The SENT Operator

Locate matches in the same sentence. For example, write a definition that locates a match for the term *Amazon* when the token *river* also occurs within the same sentence:

```
CONCEPT_RULE: (SENT, "_c{Amazon}", "river")
```

#### 4.5.1.G The SENT\_n Operator

Locate matches that occur in the specified number of sentences. For example, write a definition that locates matches for the PER concept and the term *he* within two sentences:

```
PER concept: CLASSIFIER:Obama  
CONCEPT_RULE: (SENT_2, "_c{PER}", "he")
```

#### 4.5.1.H The SENTSTART\_n Operator

Locate matches that occur within the specified number of words from the beginning of the sentence. For example, write a definition that locates matches for the term *Democratic* that occur within five words from the start of the sentence:

```
CONCEPT_RULE: (SENTSTART_5, "Democratic")
```

---

### 4.5.1.1 The SENTEND\_n Operator

Locate matches that occur within the specified number of words from the end of the sentence. For example, write a definition that locates matches on a term in the PER concept if these matches occur within five words from the end of a sentence. The following example shows how the SENT\_n, SENTSTART\_n, and SENTEND\_n qualifiers work together with a contextual operator and a classifier concept:

```
PER concept: CLASSIFIER:Obama

CONCEPT_RULE:(SENT_2, (SENTSTART_5, "Democratic"),
                (SENTEND_5, "_c{PER}"))
```

### 4.5.2 The Stemming Operator

When you add an @ symbol as a suffix to a word, you enable the expansion of the word into all of its forms. For example, if you append an @ sign to the word *book*, matches on books, booking, bookings, and so on, could be returned:

```
CONCEPT:book@
```

You can also append the @ sign followed by the letter N or the letter V to stem the word into all of its noun or verb forms, respectively. See the example below:

```
CONCEPT_RULE:(SENT, "_c{book@N}", "train@V")
```

---

**Note:** The @ symbol cannot be used in CLASSIFIER and REGEX definitions.

---

### 4.5.3 The PARA Operator

When you add the paragraph (PARA) operator, you specify that matches are located only within one paragraph. Determine the paragraph boundaries by typing one or more separator characters into the **Paragraph Separator** field in the **Project Settings - Misc** tab. When you specify more than one type of paragraph separator, use a comma (,) to identify each string as a paragraph separator. For example, you can type the following three strings to specify three different paragraph separators \n\n, \t\t, <P>.

---

You can then write one of the following rules to specify that matches can be located only in the text bounded by one or more of these separators:

```
CONCEPT_RULE: (PARA, "_c{SAS}", (OR, "statistics", "TM"))
CONCEPT_RULE: (PARA, "_c{TM}", (OR, "Enterprise Miner"))
```

#### 4.5.4 The Operators for Coreference Resolution

Coreference resolution enables you to match pronouns and other words to the canonical forms that these terms reference. (This is also known as *anaphora resolution*.) When you use coreference resolution, you can specify the canonical form of the referencing word. For example, specify *Barack, Obama,* and *President* as referring terms for the canonical form *Barack Obama*. Alternatively, choose to make *President Barack Obama* the canonical form for these terms.

For more information about coreference operators, see Section 4.8 *The Coreference Operators* on page 114.

## 4.6 Some Rule Examples

### 4.6.1 The Classifier Rules

Specify a `CLASSIFIER` rule to match one string, or dictionary entry. These rules specify a string to match in an incoming document. Unlike classifier concepts, each `CLASSIFIER` line is one `CLASSIFIER` rule.

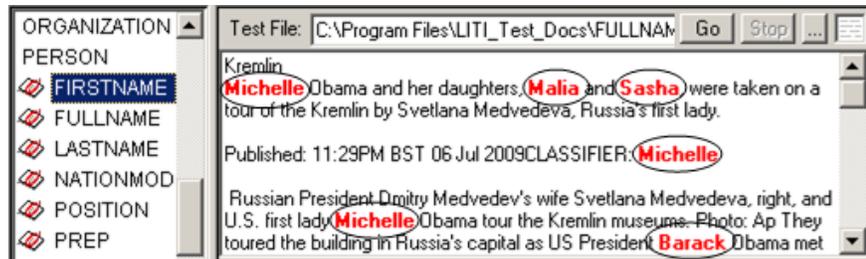
In this example, the `FIRSTNAME` concept consists of four `CLASSIFIER` rules.

*Example 4-1: Matching a Sequence of Dictionary Entries*

<b>Concept Name</b>	<b>Entry</b>
FIRSTNAME	CLASSIFIER:Sasha
	CLASSIFIER:Malia
	CLASSIFIER:Michelle
	CLASSIFIER:Barack

This **FIRSTNAME** concept matches any of the names to the right of the **CLASSIFIER** specifications in incoming texts. For example, any occurrence of Sasha, Malia, Michelle, or Barack, is a match.

Figure 4-1 *FIRSTNAME Matches in an Input Document*



**Note:** You can also specify a returned information string after a comma (,). In this case the returned information is the value for the matched concept. For more information, see *SAS Content Categorization Studio: User's Guide*.

## 4.6.2 The Sequence of Classifier Entries

Write a **CONCEPT** rule to identify related information, whether these relationships are known beforehand. For example, you might want to identify all of the lakes in the state of Michigan, but not know the names of these lakes when you write the rule. The **CONCEPT** definition specifies the ordering of **CLASSIFIER** concepts. A match occurs when matching **CLASSIFIER** strings are located in the specified order in an input document

In this example, the **FULLNAME** concept defines a relationship between the **FIRSTNAME** and **LASTNAME** concepts.

Example 4-2: *Matching a Sequence of Dictionary Entries*

Concept Name	Entry
FIRSTNAME	CLASSIFIER:Ruby CLASSIFIER:Nancy CLASSIFIER:Barack

```

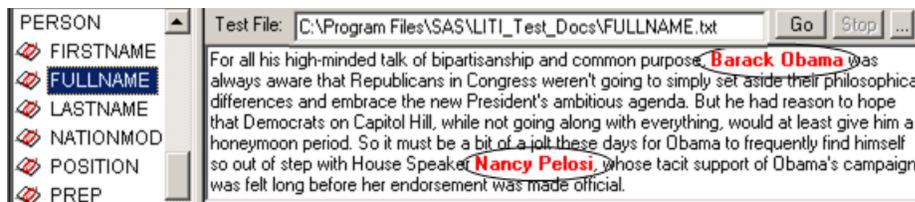
LASTNAME                CLASSIFIER:William
                        CLASSIFIER:Pelosi
                        CLASSIFIER:Obama

FULLNAME                CONCEPT: FIRSTNAME LASTNAME

```

The FULLNAME concept uses the lists of terms that are specified by the CLASSIFIER definitions in the FIRSTNAME and LASTNAME concepts. A relationship between matches on these two concepts is specified by the FULLNAME concept. For example, the terms *Nancy Pelosi* and *Barack Obama* match in an input document for both the FIRSTNAME and the LASTNAME concepts. These matches are also a match for the FULLNAME concept rule.

Figure 4-2 FULLNAME Concept Matches in an Input Document



### 4.6.3 Context Matching

Write a C\_CONCEPT rule to match text in an input document based on the context of the matches. You can also use tokens with C\_CONCEPT rules.

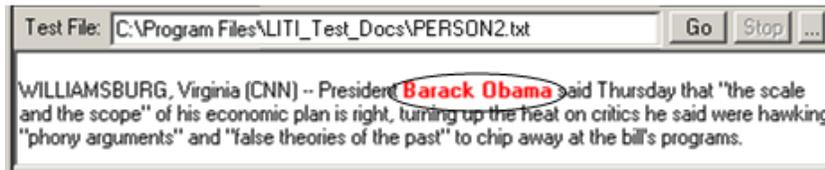
In this example, the C\_CONCEPT rule specifies a relationship between three CLASSIFIER concept rules and the token `said`.

Example 4-3: Matching within Context

Concept Name	Entry
FIRSTNAME	CLASSIFIER:Barack
LASTNAME	CLASSIFIER:Obama
TITLE	CLASSIFIER:President
PERSON	C_CONCEPT:TITLE _c{FIRSTNAME LASTNAME} said

The PERSON concept locates matches for the FIRSTNAME and LASTNAME concepts when these matches occur in the context (`_c`) specified by the curly braces (`{}`). To return a match on the rule, the preceding matches that occur are preceded by a match on the TITLE concept and followed by the token *said*. In this example, *Barack Obama* matches on the PERSON concept.

Figure 4-3 A C\_CONCEPT Match in an Input Document



#### 4.6.4 Match within Context

Write a `C_CONCEPT` definition to locate and match a word that you do not know until a match on this definition is located. However, you do know the context where these matches occur. For example, you might want to locate, and return each duplicate instance of *New Hampshire lake* in an input text.

In this example, the `C_CONCEPT` definition specifies a relationship between two matching concepts and a word beginning with an uppercase letter.

Example 4-4: Using a Reference for a Match

Concept Name	Entry
WATERBODY	CLASSIFIER:Lake
STATE	CLASSIFIER:New Hampshire
LAKES	C_CONCEPT:STATE WATERBODY <code>_c{ _cap}&gt;</code>

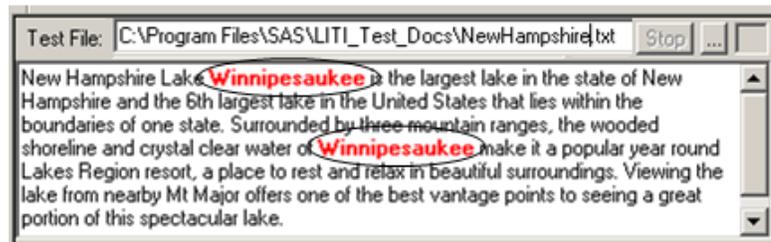
The LAKES concept specifies the context for the matched terms:

- When a match on the STATE concept is followed by a match on the WATERBODY concept, a partial match is located. For example, *New Hampshire Lake* is a partial match for this rule.
- `_c( _cap)` specifies that the matches above also appear in the context of a word that begins with an uppercase letter. In this example, a match occurs on the word *Winnepesaukee*.

- 
- By default, all of the matches in an input document are returned. When the greater than (>) symbol is specified, every instance of the matched term in the document is returned as a match regardless of the context.

In this example, two instances of *Winnepesaukee* are matched. The second match occurs because the greater than (>) symbol is specified.

Figure 4-4 C\_CONCEPT Matches in an Input Document



## 4.6.5 Eliminating Partial Matches

Specify a `NO_BREAK` rule to prevent partial matches on phrases specified in another `CLASSIFIER` definition. This rule stipulates that a match can occur only if the entire string is located in an input document.

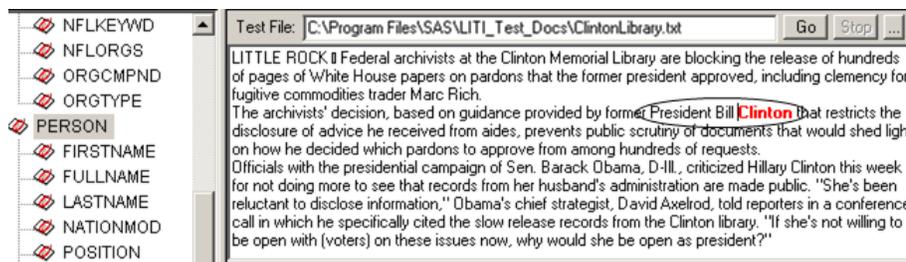
In this example, the `PERSON` concept specifies the `NO_BREAK` rule.

*Example 4-5: Excluding Spaces*

Concept Name	Entry
TITLE	CLASSIFIER:President
FIRST	CLASSIFIER:Bill
CMPND	CLASSIFIER:Clinton Memorial Library
PERSON	C_CONCEPT:TITLE FIRST _c{ _cap} NO_BREAK:_c{CMPND}

When you add the `NO_BREAK` rule to the `PERSON` concept definition, the token `Clinton` is not matched when it occurs in the phrase *Clinton Memorial Library*. However, the token `Clinton` is matched if it occurs in a string that is not specified in the `CMPND` rule. The `NO_BREAK` rule only applies to the `CMPND` rule.

*Figure 4-5 NO\_BREAK Rule Match in an Input Document*



---

## 4.6.6 Disambiguating Concepts

`REMOVE_ITEM` definitions differentiate between matches according to their context. This process of differentiation is called *disambiguation*. SAS Concept Creation enables you to specify this rule type when you refer to other concepts by writing a `REMOVE_ITEM` rule. Use this operation to eliminate a match on one rule, while returning a match on another rule.

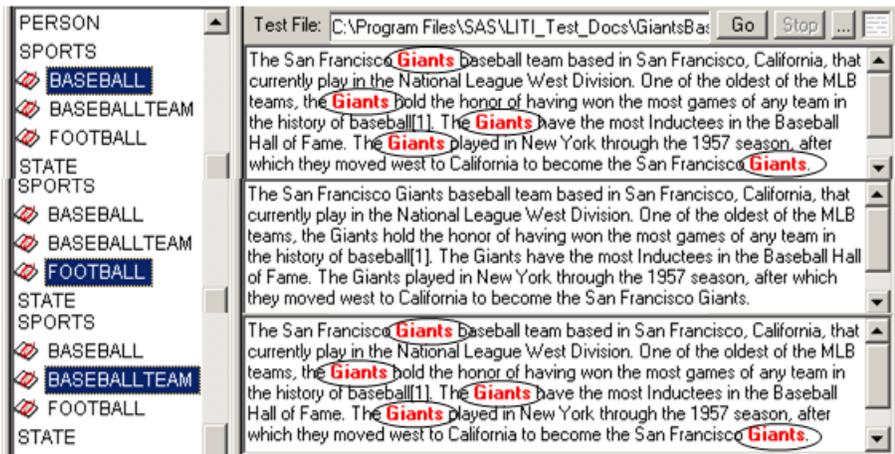
In this example, the `FOOTBALL` concept definition includes the `REMOVE_ITEM` rule to prevent Giants football documents from matching the Giants baseball concept.

*Example 4-6: Excluding Phrases*

<b>Concept Name</b>	<b>Entry</b>
BASEBALL	CLASSIFIER:Giants
FOOTBALL	CLASSIFIER:Giants REMOVE_ITEM:(ALIGNED, "_c{FOOTBALL}", "BASEBALLTEAM")
BASEBALLTEAM	C_CONCEPT:_c{BASEBALL} baseball team

Matches on the word *Giants* are returned for the `BASEBALLTEAM` concept when the token *Giants* is located in the specified context, Giants baseball team. In this case, this match is not a match for the `FOOTBALL` concept. The `REMOVE_ITEM` rule specifies that any match on both the `BASEBALLTEAM` and the `FOOTBALL` concepts only return matches for the `BASEBALLTEAM` concept.

Figure 4-6 Disambiguated Matches in Input Documents



---

## 4.6.7 Exporting Classifiers

The `CONCEPT` rule enables you to export previously unspecified classifier terms to another concept using an acronym that is specified in a concept rule. For example, specify `eLN` for last name. Alternatively, you can type the full name of the concept such as `LASTNAME`.

To write a rule using an acronym, specify this acronym in the destination rule. After an acronym is specified in a `CONCEPT` rule, other rules can specify this acronym to list the exported term.

The `CLASSIFIER` rule that specifies the export feature enables you to match incomplete terms in ways that are similar to that of the greater than symbol. For more information, see Section 4.4.8 *The > Symbol* on page 71. However, you can use only the export operation with `CLASSIFIER` rules.

In the following example, the `FULLNAME` concept specifies a `CLASSIFIER` rule that exports matches on `Sarkozy` to the `PERSON` concept that has a `CONCEPT` rule specifying `eLN`. This rule also specifies its own matching string and the context for matches.

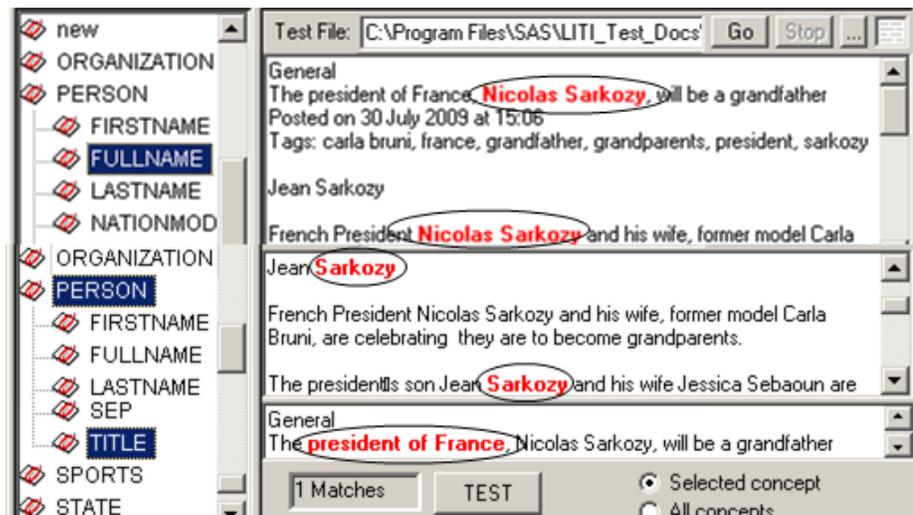
*Example 4-7: Exporting Classifiers Example 1*

<b>Concept Name</b>	<b>Entry</b>
<code>FULLNAME</code>	<code>CLASSIFIER:[export=TITLE:president of France; eLN:Sarkozy]:Nicolas Sarkozy</code>
<code>PERSON</code>	<code>CONCEPT:eLN</code>

The following matches occur in an input text that has the words *Nicolas Sarkozy* and *President of France* present somewhere in the same document:

- *President of France* is exported to, and matches, the `TITLE` concept.
- *Sarkozy* matches the `PERSON` concept. This match occurs because the acronym `eLN` is specified in the `PERSON` concept.
- *Nicolas Sarkozy* is returned as the match for the `FULLNAME` concept.

Figure 4-7 Classifier and Export Matches in Input Documents



The export feature works on an internal, per-document basis. In this example, the terms *President of France* and *Sarkozy* only match the TITLE and PERSON concepts if *Nicolas Sarkozy* is present in the input document. The exported terms do not appear in the concept definitions when these terms are exported. The concepts do not have to exist in the taxonomy in order for the export rule to work.

The COMPANY concept specifies that a match on GM is exported to the ORGANIZATION concept.

Example 4-8: Exporting Classifiers Example 2

```
COMPANY          CLASSIFIER:[export=ORGANIZATION: GM]:
                  General Motors
```

If an input text contains the string *General Motors*, the document matches the COMPANY concept. If this document also contains the word *GM*, the token *GM* is recognized as a match on the ORGANIZATION concept. However, if the word *GM* appears in a document without the term *General Motors*, *GM* is not returned as a match to the ORGANIZATION concept.

Figure 4-8 Export Rule Matches in Input Documents



---

## 4.6.8 Setting Priorities for Overlapping Matches

SAS Concept Creation enables you to override the **Priority** setting in the Data window for the selected concept. This feature works with `CONCEPT_RULE` definitions and coreference rules when you write a `PRIORITY` specification into a rule. For more information about coreference, see Section 4.8.7 *Rank Coreference Definitions and Eliminate False Positives* on page 119.

To use this feature, select **Best Matches** in the **LITI** tab of the Project Settings window.

By default, the **Priority** is set to 10 in the Data window. You can also increase the **Priority** setting in the Data window for all of the rules in one definition, or specify a `PRIORITY` in a definition. When you specify a `PRIORITY` in a rule, this setting overrides the **Priority** setting in the Data window for this rule only. The `PRIORITY` specification in a rule applies to the rule, and not to the entire definition. For this reason, any matches on this rule are prioritized over matches on any other rules in this definition, or in any other definitions

These specifications are used to increase the relative rankings between concepts. Priorities are also used to prevent matches on more than one concept. You can also use this setting to prevent matches on terms that are used in different contexts. For example, if `Roche` is specified in the `PERSON` concept and also in the `CORPORATE` concept, priorities can be used to determine the appropriate match.

The `HARBORVIEW` concept has the highest `PRIORITY` setting. Documents that match this concept, and any of the other concepts shown, are matched to the `HARBORVIEW` concept.

*Example 4-9: Setting Priorities*

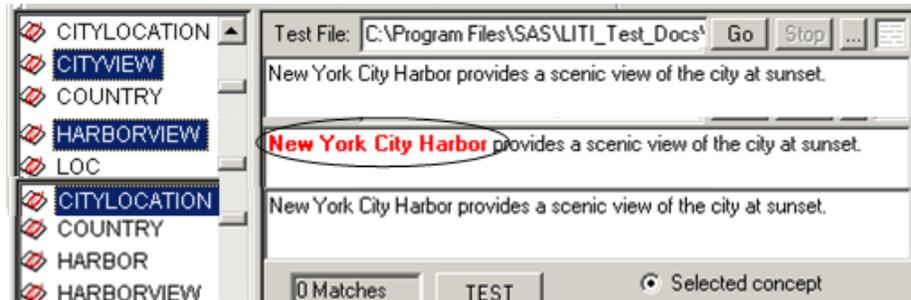
<b>Concept Name</b>	<b>Entry</b>
LOCATION	CLASSIFIER:New York
CITY	CLASSIFIER:City
HARBOR	CLASSIFIER:Harbor
CITYVIEW	C_CONCEPT:PRIORITY=20:_c{LOCATION CITY HARBOR}
HARBORVIEW	C_CONCEPT:PRIORITY=30:_c{LOCATION CITY _cap}
CITYLOCATION	C_CONCEPT:PRIORITY=25:_c{LOCATION CITY}

---

The following document is returned as a match to the HARBORVIEW concept. This is true even though *New York City Harbor* also matches the CITYVIEW concept and part of this term matches CITYLOCATION.

*New York City Harbor* provides a scenic view of the city at sunset.

Figure 4-9 A Prioritized Match in an Input Document



---

## 4.6.9 The Part-of-Speech Tags in a Definition

SAS Concept Creation enables you to use part-of-speech tags to locate matches. These tags are useful when you want to locate a wide range of matches without specifying a list of dictionary entries. Part-of-speech tags are particularly useful when you know the syntax, but not the wording of, the exact matches that you are seeking.

---

**Note:** Use the part-of-speech tags that are listed in Appendix B. Do not use the part-of-speech tags that are used with SAS Text Miner. The part-of-speech tags that are specified in your definitions are mapped to those in SAS Text Miner at the time of application.

---

A space is required before the colon (:) that precedes the part-of-speech tag. Specify a lowercase *s* in the *sep* part-of-speech tag, only.

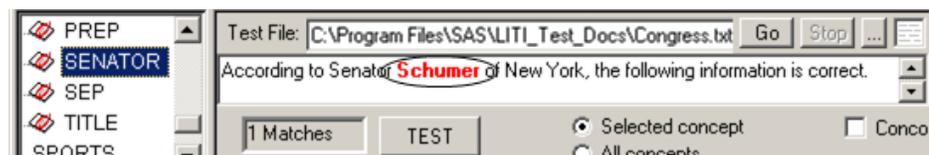
*Example 4-10: Using Part-of-Speech Tags*

Concept Name	Entry
STATE	CLASSIFIER:New York
TITLE	CLASSIFIER:Senator
SENATOR	C_CONCEPT: TITLE _c{ _cap} :Prep STATE :sep

In this example, *Schumer* is returned as a match for the SENATOR concept. This is true when a preposition (*Prep*) precedes a match on the CITY CLASSIFIER concept and a separator (*sep*) character follows this concept. See the following example:

According to Senator *Schumer* of New York, the following information is correct.

*Figure 4-10 C\_CONCEPT Rule with Part-of-speech Tag Match in an Input Document*



---

## 4.6.10 The Regular Expressions in a Definition

Specify regular expressions to locate matches based on known patterns. For example, telephone numbers, street, and e-mail addresses are all defined using recognizable patterns. When you write regular expressions, you specify a range of letters or numbers inside square braces ([ ]) to form a regular expression rule. For example, type `a-z` or `0-9`. This syntax matches any ASCII character whose value is between `a` and `z` or between `0` and `9` inclusive.

If you add a plus (+) sign after the last brace, all lowercase letters are matched. For example, you could write `REGEX: [a-z]+`.

You can also add either the % symbol or write out the word `percent`. If you perform either of these operations after you add the plus (+) symbol, all of the instances of percentages in the input document are returned.

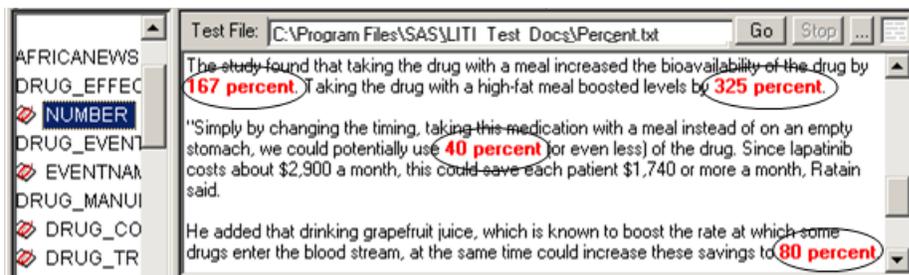
In the following example, the `NUMBER` concept has two `REGEX` rules. The two specifications for percent ensure wider definition coverage.

*Example 4-11: Specifying Regular Expressions*

Concept Name	Entry
NUMBER	REGEX: [0-9]+%
	REGEX: [0-9]+ percent

This regular expression definition specifies that numbers followed by either percentage sign match. For example, matches on both `99%`, and `50 percent` are both returned.

*Figure 4-11 REGEX Rule Matches in an Input Document*



---

**Notes:** For more information, see Appendix A *Regex Syntax and Part-of-Speech*.

---

You can also specify a returned information string after a comma (.). In this case, the returned information is the value for the matched concept.

---

### 4.6.11 The Sentence Operator in a Definition

By default, SAS Concept Creation returns matches within the entire text of an input document. Limit matches to one sentence by writing the `SENT` operator into a `CONCEPT_RULE`.

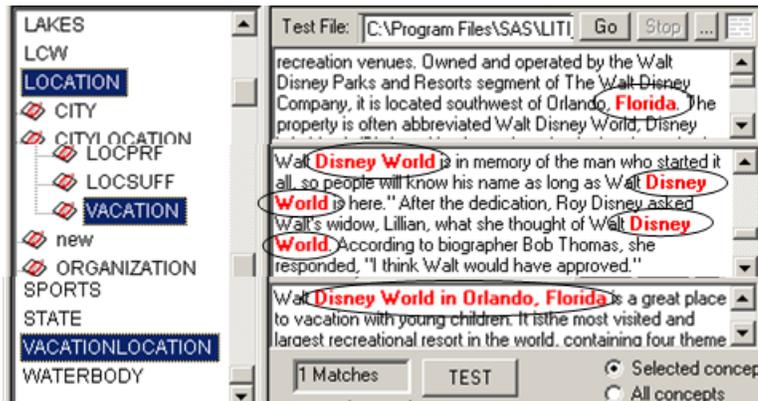
In the following example, the `VACATIONLOCATION` concept specifies that a match is returned only when all of the specified elements are located in the context of a sentence.

*Example 4-12: Specifying a Sentence Operator*

<b>Concept Name</b>	<b>Entry</b>
<code>VACATION</code>	<code>CLASSIFIER:Disney World</code>
<code>LOCATION</code>	<code>CLASSIFIER:Florida</code>
<code>VACATIONLOCATION</code>	<code>CONCEPT_RULE:(SENT, "_c{VACATION :Prep _cap :sep LOCATION}", "vacation")</code>

The `VACATIONLOCATION` definition uses the `CONCEPT_RULE` to identify a match, when all of the specified components occur within one sentence. These matches occur when a preposition follows a `VACATION` concept match, a word that begins with an uppercase letter, a separator character, and a match on the `LOCATION` concept. If this match is followed by the token `vacation`, a match is returned for the `VACATIONLOCATION` concept.

Figure 4-12 CLASSIFIER and CONCEPT\_RULE Matches In input documents



---

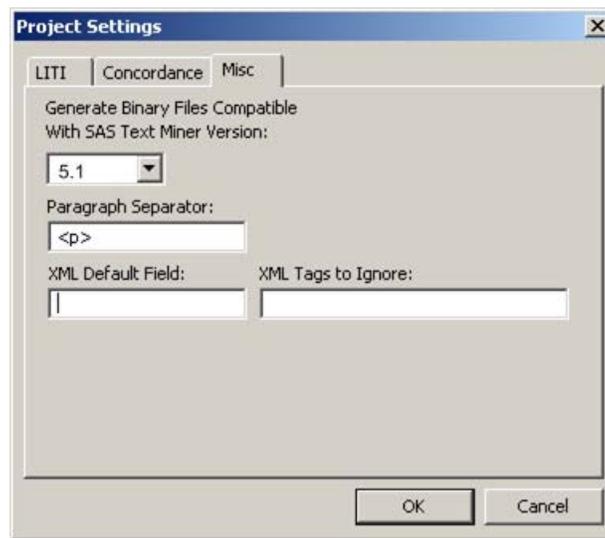
## 4.6.12 The Paragraph Operator in a Definition

By default, SAS Concept Creation looks for matches within the entire text of an input document. Limit matches to one paragraph by writing the `PARA` operator into the `CONCEPT_RULE`.

Before you specify your concept definitions, specify the paragraph separator that is used in your documents. For example, specify `<p>` for `.html` documents. If you are using multiple types of documents, list the paragraph separator for each type.

To specify the paragraph separator, complete these steps:

1. Select **Project --> Settings**. The Project Settings window appears.



2. Type the paragraph separators for your input documents into the **Paragraph Separator** field. For example, type `\n\n, \t\t, <P>`.
3. Click **OK**.

After you specify your paragraph separators, you can write the rules for each concept.

The `PARA` operator specifies that a match is returned only when all of the specified elements are located in the context of a paragraph. Each paragraph is delineated by one of these paragraph markers.

*Example 4-13: Specifying Paragraph Operators*

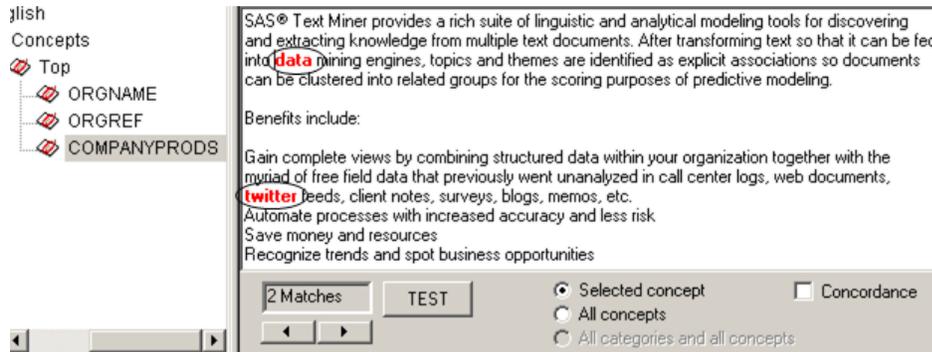
Concept Name	Entry
COMPANYPRODS	<pre>CONCEPT_RULE:(PARA, "_c{data}", (OR, "date", "engines")) CONCEPT_RULE:(PARA, "_c{twitter}", (OR, "feeds"))</pre>

The `COMPANYPRODS` definition uses two `CONCEPT_RULE` definitions to identify matches within two different paragraphs:

In the first case, a match occurs when *data* and either *date* or *engines* appear in the same paragraph.

In the second case, a match occurs when either *twitter* or *feeds* occur within the same paragraph.

*Figure 4-13 CONCEPT\_RULE and Paragraph Matches*



---

### 4.6.13 The DIST Operator in a Definition

Specify the maximum number of words between matches, instead of using the default behavior to search the entire document. The distance (`DIST_n`) operator for `CONCEPT_RULE` enables you to specify the maximum number of words that can occur between matches on the first and the last term. However, this operator does not specify the ordering of the matches.

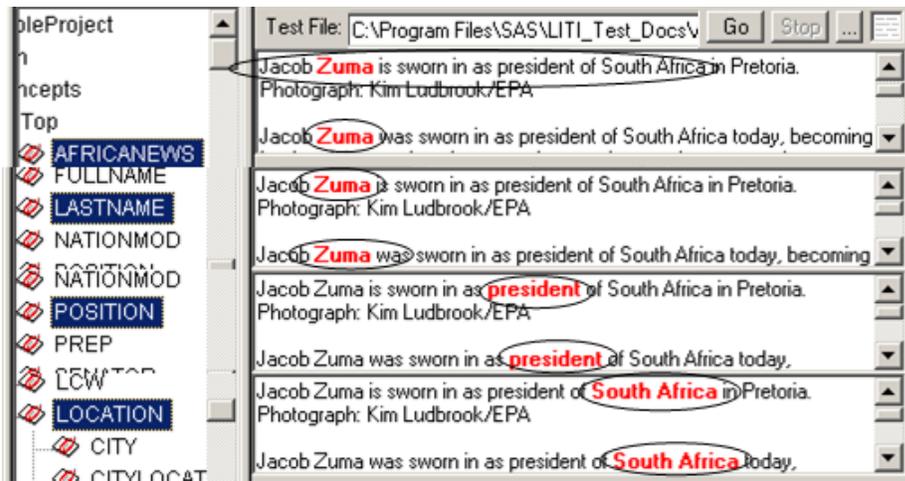
The `AFRICANEWS` definition specifies that a match is returned if there are no more than 11 words between a match on the `LASTNAME` and `LOCATION` concepts.

*Example 4-14: Specifying the DIST Operator*

<b>Concept Name</b>	<b>Entry</b>
<code>LASTNAME</code>	<code>CLASSIFIER:Zuma</code>
<code>POSITION</code>	<code>CLASSIFIER:president</code>
<code>LOCATION</code>	<code>CLASSIFIER:South Africa</code>
<code>AFRICANEWS</code>	<code>CONCEPT_RULE:PRIORITY=15:(DIST_11, "_c{LASTNAME}", "POSITION", "LOCATION")</code>

The `AFRICANEWS` concept uses the `DIST` operator to specify a distance of 11 words between the location of a match on the `LASTNAME` concept and the `LOCATION` concept. This match is returned if there is also a match on the `POSITION` concept within these 11 words. In addition, this `CONCEPT_RULE` overrides the default **Priority** setting in the Data window. If there were other rules in this definition, these rules would keep the same priority setting specified in the Data window.

Figure 4-14 CONCEPT\_RULE and CLASSIFIER Matches in Input Documents



---

## 4.6.14 The ORDDIST Operator in a Definition

The `ORDDIST_n` operator is similar to the `DIST` operator. However, the `ORDDIST` operator specifies the order and distance requirements that are necessary to return a match on the `CONCEPT_RULE` definition.

The `CONCEPT_RULE` for each `LAWFIRMS` concept places the ending curly brace (`}`) in a different location to return three different results from the same input document.

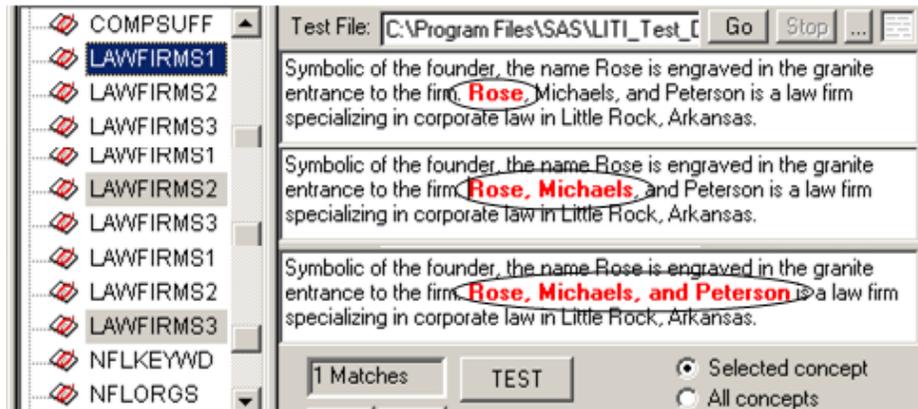
### *Example 4-15: Exporting Classifiers*

<b>Concept Name</b>	<b>Entry</b>
LCW	REGEX:[a-z]+
ORGTYP	CLASSIFIER:firm
	CLASSIFIER:firms
LAWFIRMS1	CONCEPT_RULE:(ORDDIST_15, "_c{ _cap} :sep _cap :sep and _cap", "LCW", "ORGTYP")

This `CONCEPT_RULE` states that the following instances return a match if the matches occur in the specified order and within a distance of 15 words. A word begins with an uppercase letter and is followed by a separator character and an uppercase letter. This match is followed by a separator character, the token `and`, and another word beginning with an uppercase letter. The match is not returned unless the `LCW` `REGEX` rule is also matched and a match on the `ORGTYP` concept also occurs within 15 words.

When the closing curly brace (`}`) is moved for the `LAWFIRMS2` and `LAWFIRMS3` concepts, the following matches are returned.

Figure 4-15 CONCEPT\_RULE Matches in Input Documents



You can also change the default **Priority** setting of 10 in the Data window for any of the three concept definitions shown above.

Display 4-1 Project Settings - LITI Settings



Use the Project Settings to affect how matches are returned:

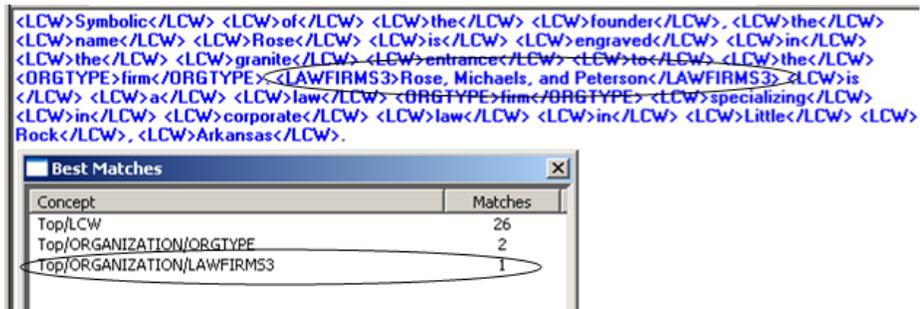
- Select **All matches** and all of the matches for LAWFIRMS1, LAWFIRMS2, and LAWFIRMS3 above are returned. In this case,

because the greater than (>) symbol does not end any of the CONCEPT\_RULE definitions, only one match is returned for each concept.

- Select **Longest** and a match on LAWFIRMS3, only, is returned.
- Select **Best** and a match LAWFIRMS3 is returned. This is true unless you specify a higher priority in either the Data window or within a concept definition.

See the example shown below that applies to both the **Longest** and **Best** selections:

Figure 4-16 Best Matches Screen



- Select **Return all identical matches**, if either **Longest** or **Best** is matched, and all of the instances with the same priority or length are returned.
- The **Remove duplicate facts** operation does not apply. No facts can be specified for CONCEPT\_RULE definitions.

---

## 4.7 Locating Facts

### 4.7.1 Overview of Facts

Facts, or predicates, refer to terms that match at least two concepts. Facts consist of at least two arguments. For example, *Harry Truman was president of the United States* is a fact based on three arguments. These arguments are defined by the following concepts NAME, TITLE, and COUNTRY. The following matches *Harry Truman*, *president*, and *United States* are returned to these concepts. By specifying this type of rule, you also locate similar matches in input documents without rewriting your rules.

Both `SEQUENCE` and `PREDICATE_RULES` extract facts. `SEQUENCE` rules specify the order of the matches. `PREDICATE_RULES` use Boolean operators, but do not specify the ordering of any matches. For more information, see Section 4.7.2 *The Predicate Sequence Example* on page 107 and Section 4.7.3 *The Predicate Examples* on page 109.

### 4.7.2 The Predicate Sequence Example

Identify previously unknown relationships, otherwise known as facts or events, in input documents. Predicate sequence, or `SEQUENCE`, rules extract the meaningful relationships between matched concepts and tokens. For example, identify the names and positions that various managers hold within a company. Locate this information even when these relationships are unknown to you, or when the concepts do not directly follow one another.

Predicates are also defined as facts or events. The terms are interchangeable. Facts are always defined by at least two concepts or tokens and one or more parts of speech. The term *sequence* is used to specify the necessary ordering of the concepts and semantic terms that define these facts.

When you specify a predicate sequence definition, you define not only the concepts, but also the arguments that are used with these concepts. Use this rule to also specify the sequence of these entities and any appropriate parts of speech.

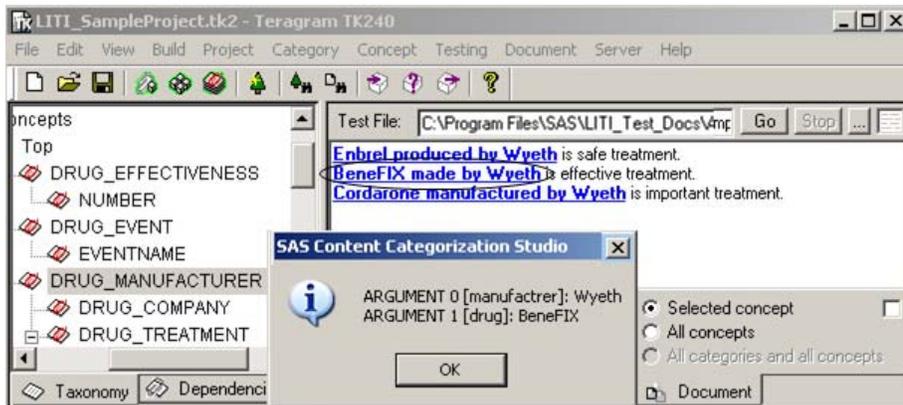
Example 4-16: Writing a Predicate Sequence Definition

Concept Name	Entry
DRUG_COMPANY	CLASSIFIER:Wyeth
DRUG_MANUFACTURER	SEQUENCE: (drug,manufacturer): _drug{ _cap } _w _w _manufacturer { DRUG_COMPANY } _w _w treatment

This SEQUENCE rule takes two arguments, drug and manufacturer. To locate the \_drug predicate, locate a word that begins with an uppercase letter that is followed by two tokens. To match the \_drug predicate, locate the DRUG\_COMPANY concept followed by two tokens and the word treatment. However, only the matches within and between the beginning and ending curly braces ({} ) are returned as a match for this concept.

For example, the fact *BeneFIX produced by Wyeth* is returned as a match to the DRUG\_MANUFACTURER SEQUENCE concept along with the matches on the arguments for this fact. You can see the fact matches in the Document window for this testing document. You can also click on one of the returned facts to access a SAS Content Creation status screen. This screen lists the matching arguments for the selected fact.

Figure 4-17 Argument Matches in an Input Document



---

### 4.7.3 The Predicate Examples

Like `SEQUENCE` rules, `PREDICATE_RULES` locate facts and their supporting arguments. Unlike `SEQUENCE` rules, `PREDICATE_RULES` do not specify the matching order. Instead, `PREDICATE_RULES` use Boolean operators to increase the matching precision within the document. For more information, see Section 4.5 *The Operators* on page 79.

Like the preceding `SEQUENCE` rule, this `PREDICATE_RULE` defines two arguments, `drug` and `manufacturer`. However, the `DRUG_MANUFACTURER` `PREDICATE_RULE` uses the `DIST` operator. This operator specifies that a match is returned when the `DrugName` concept is located within 20 words of a match on the `DRUG_COMPANY` concept.

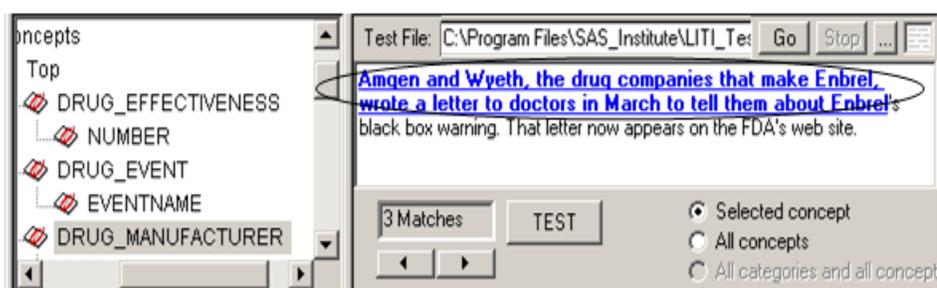
*Example 4-17: Viewing a PREDICATE\_RULE*

Concept Name	Entry
DrugName	CLASSIFIER:Enbrel
DRUG_COMPANY	CLASSIFIER:Amgen CLASSIFIER:Wyeth
DRUG_MANUFACTURER	PREDICATE_RULE: (drug,manufacturer): (DIST_20, "_drug{ DrugName }", "_manufacturer{ DRUG_COMPANY }", "make")

This `PREDICATE_RULE` defines two arguments, `drug` and `manufacturer`. Inside the parentheses that follow each argument is the concept that identifies a match. The `DIST` operator specifies that matches on the `DrugName` concept can occur within 20 words of a match on the `DRUG_COMPANY` concept. In addition, a match on the `DRUG_MANUFACTURER` concept only occurs when the token `make` is located. Although no other tokens are specified for this `PREDICATE_RULE`, all of the words located between matches on the concepts `DrugName` and `DRUG_COMPANY` are returned as a matching phrase. However, because a `PREDICATE_RULE` is specified and not a `SEQUENCE` rule, these matches can occur in any order.

For `PREDICATE_RULES`, like other definitions, multiple matches can occur in one document, and multiple facts can be returned.

Figure 4-18 PREDICATE\_RULE Match in an Input Document



The results shown above are returned when the default setting, **All matches**, is selected under the **Overlapping Concept Matches** heading in the Project Settings - LITI dialog box.

Click  and  in the Document window to see each of the following matches:

Amgen and Wyeth, the drug companies that make Enbrel

This fact matches the word *Wyeth* as a token. It is not a match on the *DrugName* concept.

Wyeth, the drug companies that make Enbrel

This is the shortest of the two matches that begin with a match on *wyeth* in the *DRUG\_COMPANY* concept and end with *Enbrel* as a match on the *DrugName* concept. Also see the following bulleted point.

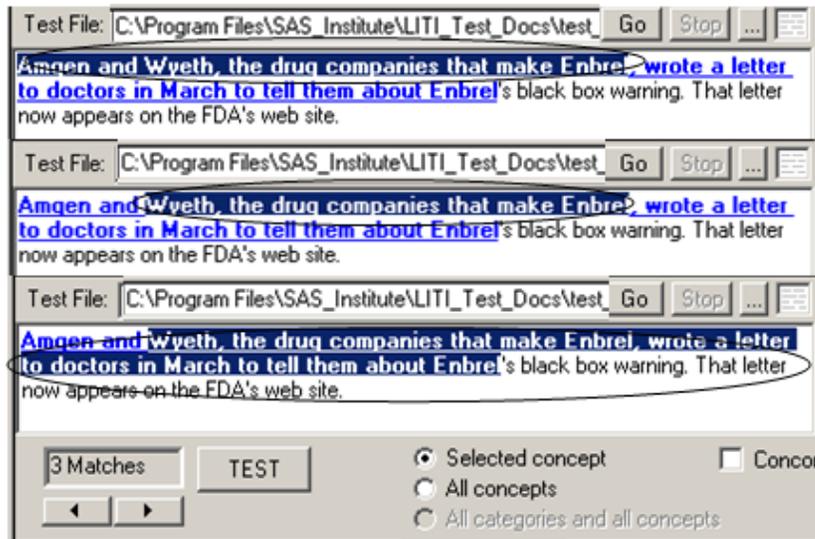
Wyeth, the drug companies that make Enbrel, wrote a letter to doctors in March to tell them about Enbrel

This is the longest of the two matches that begin with a match on *wyeth* in the *DRUG\_COMPANY* concept and end with *Enbrel* as a match on the *DrugName* concept. In this case, the first instance of *Enbrel* is matched as a token and not as a match on the *DrugName* concept. Also see the bulleted point above.

This match is returned when you select **Longest** under the **Overlapping Concept Matches** heading in the Project Settings - LITI dialog box.

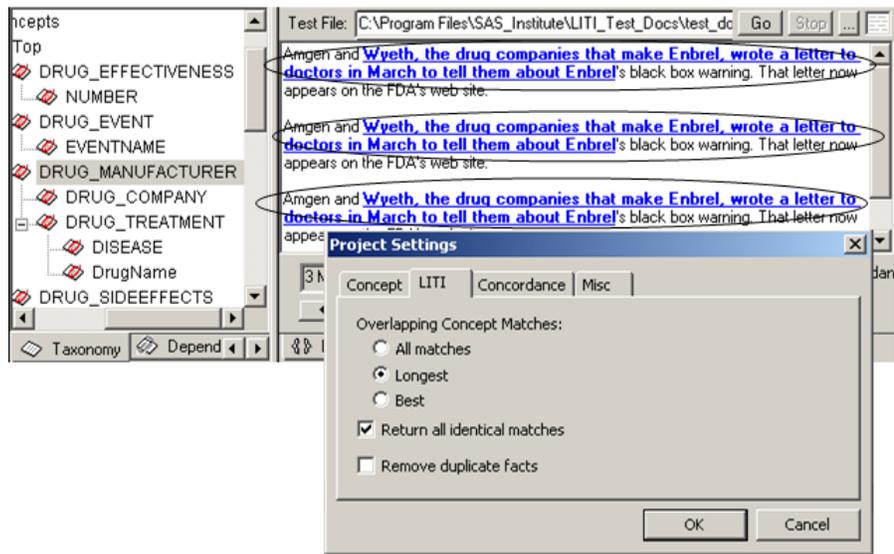
These results are also returned when you select **Best**. This statement is true unless you set a **Priority** specification in the **Definition** tab or overwrite the default setting of 10 in the Data window for this concept.

Figure 4-19 PREDICATE\_RULE Matches in Input Documents



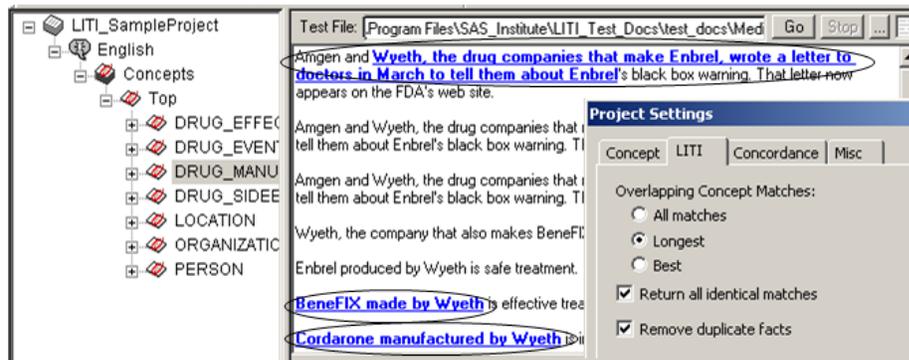
To return all of the instances of the longest fact matches, select **Return all identical matches** in the Project Settings - LITI dialog box. This operation can only be selected if you have also selected either **Longest** or **Best** under the **Overlapping Concept Matches** heading.

Figure 4-20 Several Instances of a Match in an Input Document



- In the figure below, **Remove duplicate facts** is added to the selections in the figure above. New text is added to the testing document to illustrate the functionality of these interrelated settings. Each instance of a match that is the longest for any of the overlapping matches, but not a duplicate fact, is returned as a match to the selected concept.

Figure 4-21 Longest, Unique Matches in an Input Document



All three of these facts are highlighted, and initially appear as a single match to the PREDICATE\_RULE definition for the DRUG\_MANUFACTURER concept. However, there are two sets of arguments, because there are two matches on the DRUG\_COMPANY concept and one match on the DrugName concept. It is these matches that define the beginning and end of each fact.

Figure 4-22 Facts and Arguments in an Input Document



---

## 4.8 The Coreference Operators

### 4.8.1 Overview of Coreference

Use coreference operators to write rules that return the canonical form of a word along with the referring term. Coreference operators are often used with pronouns, or other words that are called *referring terms*. (This is also known as *anaphora resolution*.) The canonical form of a word can be any term that you choose. For example, return either *Barack Obama* or *President Barack Obama* as a match for each instance of the referring term *Barack* in an input document. Another alternative is to choose to return *President Barack Obama* as the canonical form for each match on the pronoun *he*.

When the tested document is displayed in the **Document** tab, both the canonical word form and the matching term are highlighted.

Use the coreference operator (`_ref`) with a `CONCEPT`, `C_CONCEPT`, or a `CONCEPT_RULE` rule. If you want to use a coreference qualifier in a `CLASSIFIER` rule, use `_coref` instead of `_ref`.

---

**Note:** The **Overlapping Concept Matches** selections in the **LITI** tab of the Project Settings window do not affect matches made by the export, forward, and preceding operators.

---

### 4.8.2 How to Use the Coreference Operator

Use the coreference operator (`_ref`) to link a matched string with its canonical form in an input document.

```
C_CONCEPT:{Jim Goodnight} said _ref{he}
```

In the example above, the canonical form *Jim Goodnight* is returned each time the matching term, *he* is located. This is true when the phrase *Jim Goodnight said he* is located in the text.

The `_c` operator is used in a `C_CONCEPT` rule that specifies the canonical form for the coreference specified by the `_ref` operator.

Example 4-18: C\_CONCEPT Rule with the \_ref Operator

Concept Name	Entry
PLNOUNGROUP	CLASSIFIER:Democratic leaders
PERSON	C_CONCEPT:_c{PLNOUNGROUP} said _ref{they}

When this definition is matched in an input document, a match on the referring term that follows the `_ref` operator returns the canonical form. The canonical form is specified in the bracketed term that follows the context operator (`_c`). This form is identified in the PLNOUNGROUP concept. In this example, the word that *they* references its specified canonical form *Democratic leaders*.

Figure 4-23 `_ref` Match in an Input Document



In this example, *Democratic leaders* and *they* are returned as matches in this input document. However, if the document contained other instances of the word *they*, these instances are not matched. You can see these matches in the Document window for this testing document.

### 4.8.3 How to Use the `_ref` Operator with the `>` Symbol

The greater than symbol (`>`) locates multiple instances of a match specified by the bracketed (`{}`) coreference operator (`_ref`) in an input document. For example, you might want to return the canonical form for each matched instance of a first name. In this case, you could specify a rule that identifies any references to *Jim* as a reference to *Jim Goodnight CEO of SAS Institute*. For more information, see Section 4.4.8 *The `>` Symbol* on page 71.

---

## 4.8.4 How to Use the `_ref` Operator with the Forward or Backward Symbols

### 4.8.4.A Limiting Matches to Those That Follow or Precede a Coreference Match

Use the forward (`_F`) and the preceding (`_P`) symbols to restrict coreference matches in an input document. When you specify these operators, only the matches that follow or precede the match for the rule, respectively, are returned.

Use these symbols when you want to return all of the matches instead of the one match that follows the rule (`coref` operator alone). Unlike the greater than (`>`) symbol, all of the returned matches can occur only before or after the coreference rule match.

### 4.8.4.B Matching with the Forward Symbol

Use the forward symbol (`_F`) to return all of the matches that follow a coreference rule match.

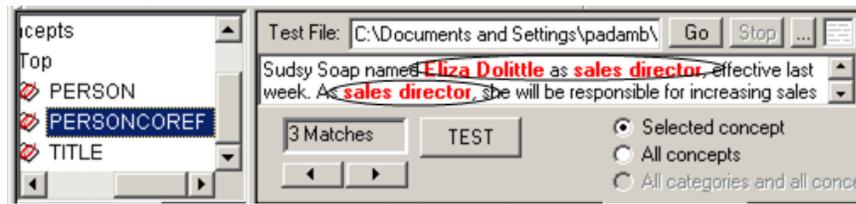
The example below shows a concept with a concept rule with a forward symbol. The rule specifies that all of the instances of matches on the coreference term that follow the coreference match are returned as matches. (Any matches that precede the match on the coreference term are not returned.)

*Example 4-19: Two C\_CONCEPT Rules with the `_F` Symbol*

<b>Concept Name</b>	<b>Entry</b>
PERSON	CLASSIFIER:Eliza Dolittle
TITLE	CLASSIFIER:sales director
PERSONCOREF	C_CONCEPT:_c{PERSON} as _ref{TITLE}_F

In this example, a match on the term *Eliza Dolittle* as *sales director* matches. Instances of the term *sales director* that follow are also returned as matches.

Figure 4-24 *\_ref* and Forward Symbol Matches



#### 4.8.4.C Matching with the Preceding Symbol

Use the preceding symbol (*\_P*) to return matches on all instances of a coreference match that occur before the coreference rule match.

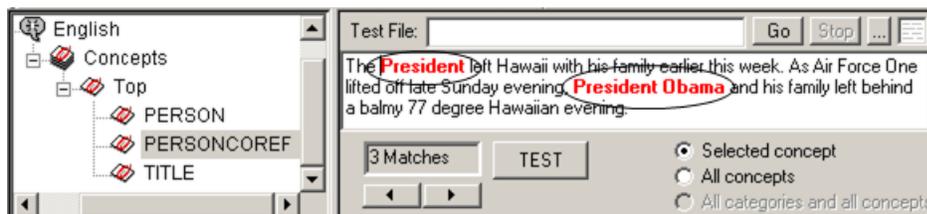
The example above shows a concept with a rule that specifies a preceding symbol. All instances of matches on the TITLE concept that are immediately followed by a match on the PERSON concept are returned as matches. (Any matches that follow the match on the coreference term are not returned.)

Example 4-20: Two C\_CONCEPT Rules with the *\_P* Symbol

Concept Name	Entry
PERSON	CLASSIFIER:Obama
TITLE	CLASSIFIER:President
PERSONCOREF	C_CONCEPT:_ref{TITLE}_P _c{PERSON}

In the example above, all instances of a match on the TITLE concept that precede a match on the TITLE and PERSON concepts are matched in an input document.

Figure 4-25 Matches on a Rule with the Preceding Operator



---

## 4.8.5 Coreference in a Classifier Definition Example

You can use the coreference operator (`coref`) to link a match in a coreference definition to its canonical form. For example, you might want to return *Barack Obama* for a match on any instance of the word *president* in an input document. The `coref` qualifier is used with classifier definitions, only.

The example above shows a classifier definition that links matches on the `coref` qualifier to its canonical form.

*Example 4-21: A Classifier Concept with a Coreference Qualifier*

Concept Name	Entry
FULLNAME	CLASSIFIER:[coref=Clinton,William Clinton;TITLE:President]:Bill Clinton

In the example above, if the canonical term *Bill Clinton* is matched once in an input document, all instances of matches on the `coref` qualifier terms also return matches. In this example, *Clinton*, *William Clinton*, and *President* all return matches. The canonical form for each matched term is *Bill Clinton*.

## 4.8.6 Assigning New Concept Names to Coreference Matches

You can assign a new concept name for a match on a term specified by the `_ref` operator. In this case, any instances of this match are output in SAS Content Categorization Server as a match on this new concept. You can also write a rule that specifies that a match is assigned to an existing concept. For example, you could assign matches on the names of an organization to an existing `CLASSIFIER` definition. In both cases, any matches on the complete definition are returned in the specified canonical form.

Specify a new, or an existing, concept name in square brackets (`[]`) that are preceded by the `_ref` operator. For example, specify `_ref [COMPANY]`.

In the example above, if a sequence of two or more words that begins with an uppercase letter is followed by *Inc.*, a match is returned for the `ORGREF` concept. A sequence of two words that begin with uppercase characters is returned as a match for the concept `ORGNAME`. The canonical form is returned as a match for the `ORGREF` concept.

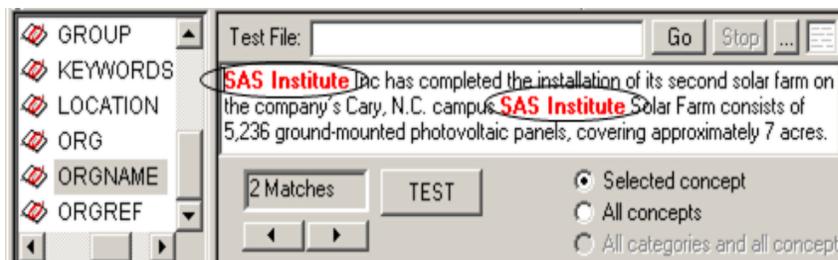
---

Example 4-22: Assigning a New Concept Name to a Coreference Match

Concept Name	Entry
ORGRAME	CLASSIFIER:SAS Institute Inc.
ORGRF	CONCEPT:_ref[ORGRAME]} {_ref ( _cap)> _cap}> Inc.

In the example above, a match on the ORGRAME concept is returned when there is a match on the remainder of the ORGRF rule. For example:

Figure 4-26 Match Returned to Another Concept



#### 4.8.7 Rank Coreference Definitions and Eliminate False Positives

You can use the `PRIORITY` specification to make matches on one coreference rule rank higher than other rules. Specify a priority to rank matches on the concept that uses coreference higher than other matched concepts.

You can choose to specify a priority for a concept match that uses the `_ref` operator with the export symbol. You can also use the `PRIORITY` specification to eliminate false positives. For more information about priorities, see Section 4.6.8 *Setting Priorities for Overlapping Matches* on page 94.

In this example, if *Samuel A. Alito Jr.* is present once in the document, every match on *Alito* returns his full name. The canonical form is *Samuel A. Alito Jr.* and the referring term is *Alito*.

Example 4-23: `C_CONCEPT` Rule with the Export Symbol

Concept Name	Entry
FIRST	CLASSIFIER:Samuel

```

INITIAL                CLASSIFIER:A.
PERSUFFIX             CLASSIFIER:Jr.

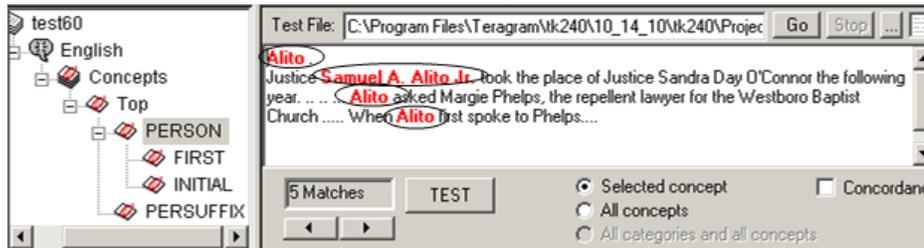
PERSON concept:      CONCEPT:PRIORITY=30:FIRST INITIAL
                    _ref{ _cap }> PERSUFFIX

```

In the example above, all instances of *Alito* are matched in an input document when all of the following conditions are met. A match on a first name listed in the FIRST classifier concept is located. This match is followed by a match on an initial specified in the INITIAL concept. When a word beginning with an uppercase letter follows this match, it is the coreference that is matched by all instances that occur in the document. Finally, a match on the PERSUFFIX concept is located.

In the example shown below, all instances of *Alito* are returned as a match. The PERSON concept also has a priority setting of 30. This means that matches on the PERSON concept rank higher than the matches that are also returned to the FIRST and INITIAL definitions.

Figure 4-27 *\_ref* and Export Symbol Matches



## 4.9 XML Fields in Rules

### 4.9.1 Overview of XML Field Matching

If the input is a valid XML document, SAS Concept Creation enables you to write rules that restrict matches to the fields that you specify. These are the ways to process XML documents:

- 
1. Specify default fields in the rules.
  2. Specify field names in the rules.
  3. Combine both operations.

By default, text is extracted from all of the fields before matching takes place. If you want to restrict matching to specific fields, you can specify these fields in the **XML Default Field** of the **Misc** tab in the Project Settings interface.

You can specify one XML field with the `CLASSIFIER`, `CONCEPT`, `C_CONCEPT`, `SEQUENCE`, `NO_BREAK`, and `REGEX` rules. Specify the field name at the beginning of the pattern to be matched. For example, specify the `body` field as the location where all matches occur.

---

**Note:** Matches are returned only if the matches are located within, and not across, fields.

---

## 4.9.2 The SEQUENCE Rule with an XML Field Example

When you write a `SEQUENCE` rule, all of the individual tokens or concepts are matched. These matches occur if all of the tokens and concepts are present within the specified field. `SEQUENCE` rules do not enable matching across fields.

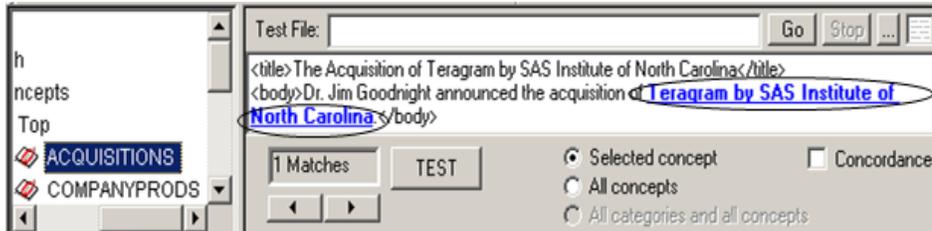
The XML field is preceded by an underscore (`_`) and the concepts to be matched follow. In the `SEQUENCE` rule example above, there are three arguments. A match occurs when each of these arguments is matched in the `body` field of an input XML document.

*Example 4-24: Assigning a New Concept Name to a Coreference Match*

<b>Concept Name</b>	<b>Entry</b>
ORG	CLASSIFIER:SAS Institute CLASSIFIER:Teragram
LOCATION	CLASSIFIER:North Carolina
ACQUISITIONS	SEQUENCE:(org1,org2,loc):_body: acquisition of _org1{ ORG } by _org2{ ORG } of _loc{LOCATION}

A match for the ACQUISITIONS concept occurs when the term *acquisition of* occurs followed by two matches on the ORG concept separated by the word *by*. This match is complete when it is followed by a match on the LOCATION concept and all of these matches occur in the `body` field.

Figure 4-28 Match Located in an XML Field



### 4.9.3 Matching More than One XML Field

If you choose to use a `PREDICATE_RULE`, `CONCEPT_RULE`, or a `REMOVE_ITEM` definition, you can specify a separate field for each argument.

Each XML field is preceded by an underscore (`_`). For example, `_title` and `_p`. The specified matches are enclosed in quotation marks ("`"`"). See the following example:

Example 4-25: Matching Two XML Fields

Concept Name	Entry
GROUP	CLASSIFIER:Trio
ORGNAME	CLASSIFIER:ABC Inc. CLASSIFIER:Sue Ann Kahn
COMPOSERS	PREDICATE_RULE:(inst1,org2):(AND, _title:"_inst1{ GROUP }",_p:"by _org2{ ORGNAME }")

A match for the COMPOSERS concept occurs when there is a match in the `title` field on the GROUP concept. The match is complete when there is also match on the `p` (paragraph) field on the word *by* followed by a match on the ORGNAME concept. (The field name is preceded by an underscore [`_`])

---

## 4.10 Writing Multiple Rules for One Definition

Write multiple rules for each concept. This feature increases the recall of your definitions by enabling you to locate more matches as well as matches based on different specifications.

For example, add the `SEQUENCE` rule shown in Example 4-16 on page 108 to the definition of the `DRUG_MANUFACTURER` concept to locate matches in documents that might not otherwise match.

## 4.11 Troubleshooting Your Rules

If you do not obtain the results that you expect, or if SAS Concept Creation returns syntax error messages, troubleshoot your rules.

To troubleshoot your rules, use the following list:

- **Case sensitivity:** Have you specified your rules to match the upper- and lowercase words that you want to match?
- **sep part-of-speech:** Did you remember to specify `sep` beginning with a lowercase `s`?
- **Project Settings:** Are these settings returning the best results?
- **Rule type:** Did you specify the correct rule type using all uppercase letters?
- **Spaces:** Did you remember to use spaces before the colon (`:`) that precedes part-of-speech tags?
- **Curly braces** (`{}`): Did you surround the term that you want to return with curly braces?
- **Square braces** (`[]`): Did you surround the new, or other, concept to be matched with square braces when you wrote a coreference rule?
- **Syntax:** Have you checked the rule syntax using the **Syntax Check** button in the **Definition** tab before compiling your concepts? Is this syntax appropriate for the results that you are trying to return, or is there a better syntax or rule type?



---

## Part 2: Testing

---

- Chapter 5: *Assembling Testing Sets on page 127*
- Chapter 6: *Testing the Concept Definitions on page 143*



---

# Chapter: 5

## Assembling Testing Sets

---

- *Overview of Assembling Testing Sets*
- *Creating Testing Folders*
- *Collecting Test Files*
- *Import Test Files*
- *Delete Testing Files*

### 5.1 Overview of Assembling Testing Sets

You gather groups of documents, or testing sets, for the purposes of testing the concept definitions that you develop in SAS Concept Creation for SAS Text Miner (SAS Concept Creation). These testing documents are used to see whether you are obtaining the results that you expect. Test your concepts before they are applied as custom entities by SAS Text Miner.

To set up a directory of testing documents, choose texts for each concept. These are the documents that you expect to match the definition for that concept. Place each set of these texts into a testing folder. Create one folder for each taxonomy node.

Testing documents help to determine whether, and why, a concept definition should be changed so that the rule correctly extracts results. For this reason, the test files that together comprise the testing set, or sets, of documents are integral to developing an accurate SAS Concept Creation project. The process of testing and refining rules can be used reiteratively until you obtain a required set of definitions.

After you test the testing directory, set up a central repository that is one folder of testing documents. Place documents that are similar to the real world texts that you plan to input to SAS Text Miner into this folder. These documents are not matched to individual concepts. For this reason, the central repository is a large group of documents that test the entire taxonomy. This folder can also

---

contain test documents that should fail but might not. For example, if you want to match *capital cities*, you might include texts that have the word *capital* meaning uppercase letter in your failing folder.

Before you test your concept definitions, use the directions in this chapter to develop each of the types of testing directories that you want to use. An overview of the process detailed in this chapter is provided below:

1. Create the directory of testing folders for individual concepts that matches the taxonomy.
2. Collect 5 - 10 documents that you expect to match each concept.
3. Place these testing documents into the folders that you created.
4. Set the paths to these files.

You can also automate some of these steps. For example, you can create a top level testing folder and use the **Create Folders** check box and the **Propagate** button in the Data window. These operations simultaneously create testing subdirectories and set the paths to these directories. For more information, see Section 5.2.1 *Create a Testing Directory While You Set Paths* below. Read this chapter before you decide how to create your testing folders.

## 5.2 Creating Testing Folders

### 5.2.1 Create a Testing Directory While You Set Paths

Use SAS Concept Creation to automatically develop the testing directory while setting the testing paths to these folders. This operation saves time and ensures that an exact replication of the taxonomy displayed in the Taxonomy window is copied for the testing documents.

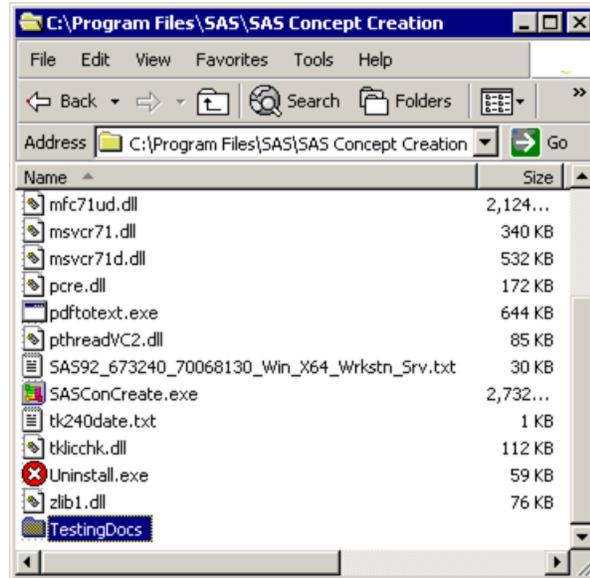
---

**Hint:** If you rename a concept, you might also want to change the name of the testing folder.

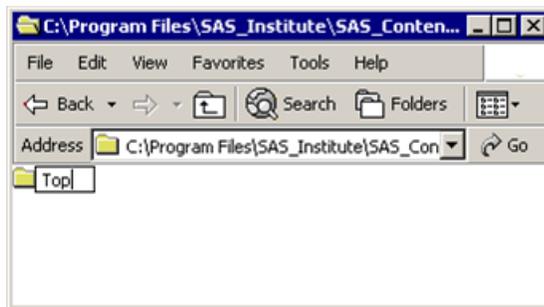
---

To define the testing taxonomy while simultaneously setting the testing paths, complete these steps:

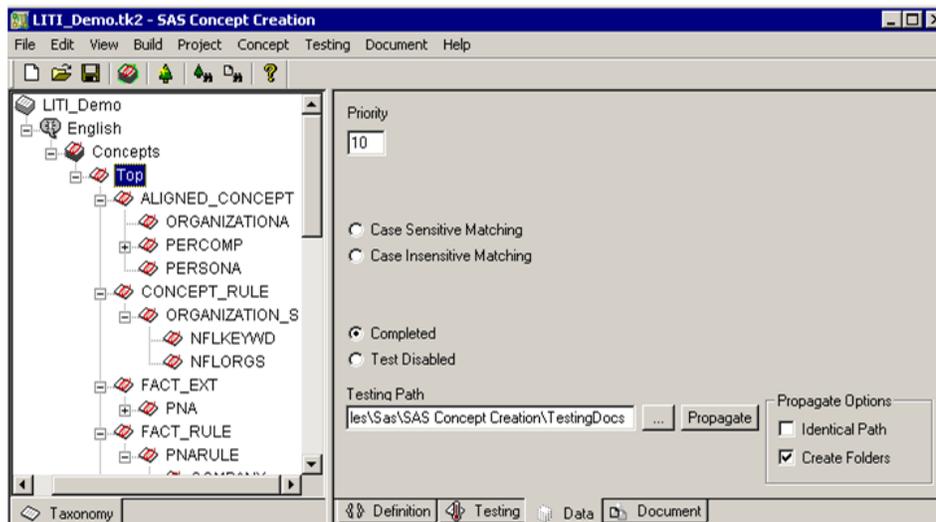
1. Access the folder for your project and create a new file for the testing documents. Name this folder. For example, type `TestingDocs` into the name space for this directory.



2. Double-click the testing folder and create a new folder named *Top* to match the `TOP` folder in the Taxonomy window. This folder is used to automatically propagate the testing paths to each of the concepts and their children in your taxonomy.



- 
3. Select the `TOP` folder in the Taxonomy window.



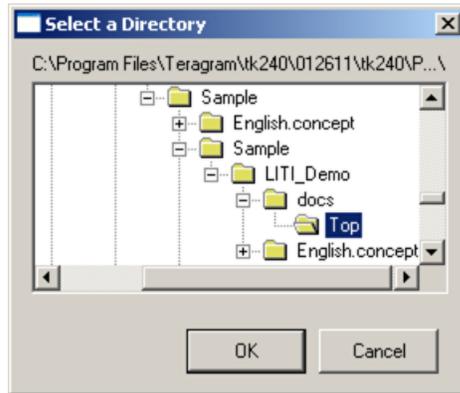
---

**Note:** If you click another node, SAS Concept Creation creates only subdirectories for the selected concept node.

---

4. Select **Create Folders** under the **Propagate Options** heading in the Data window.

- 
- Click  to the right of the **Testing Path** field and the Select a Directory window appears.



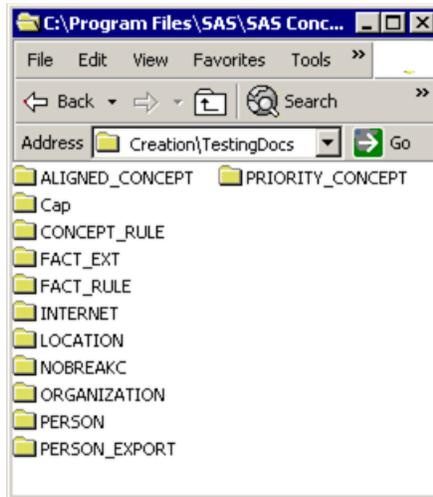
- Select the `TOP` directory where SAS Concept Creation creates the testing taxonomy.
- Click **OK**.
- Click **Propagate** in the **Data** tab. A SAS Concept Creation confirmation window appears.



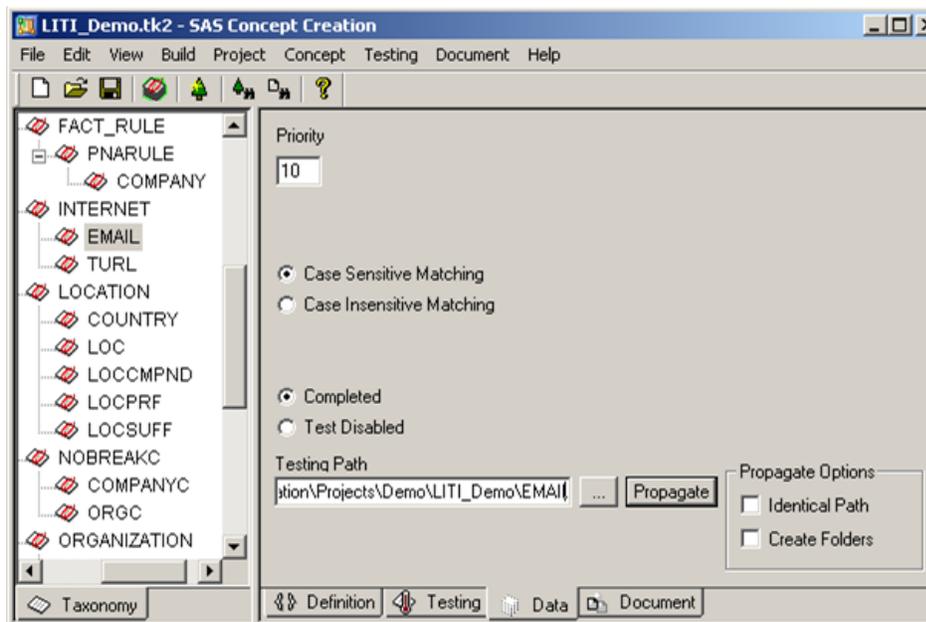
- Click **OK**.

---

A directory structure that is identical to the taxonomy is created inside the `Top` folder.



- 
10. Click some of the nodes in the Taxonomy window to see that each **Testing Path** field displays the path to the matching testing directory. See the following example:



Unless each folder in the testing directory is populated with your testing documents, you cannot test your concepts. You can also choose to manually add additional documents to your testing folders.

---

## 5.2.2 Create and Set a Path to the Central Repository

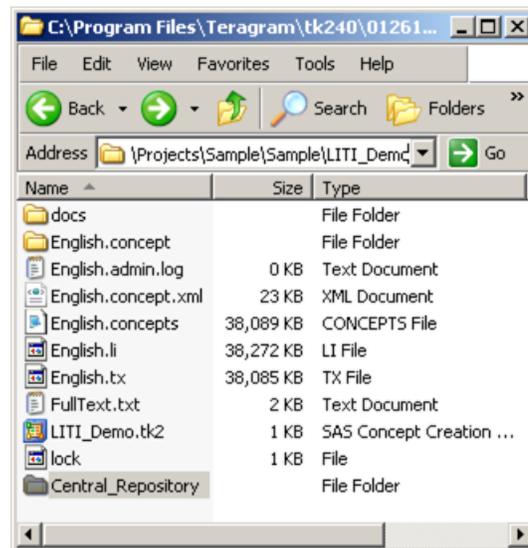
A central repository of testing documents contains a set of texts that are not selected to match individual concepts. For this reason, when you test the central repository, you gain a realistic approximation of the results that you might obtain for real-world documents.

Use a central repository of testing documents for the following purposes:

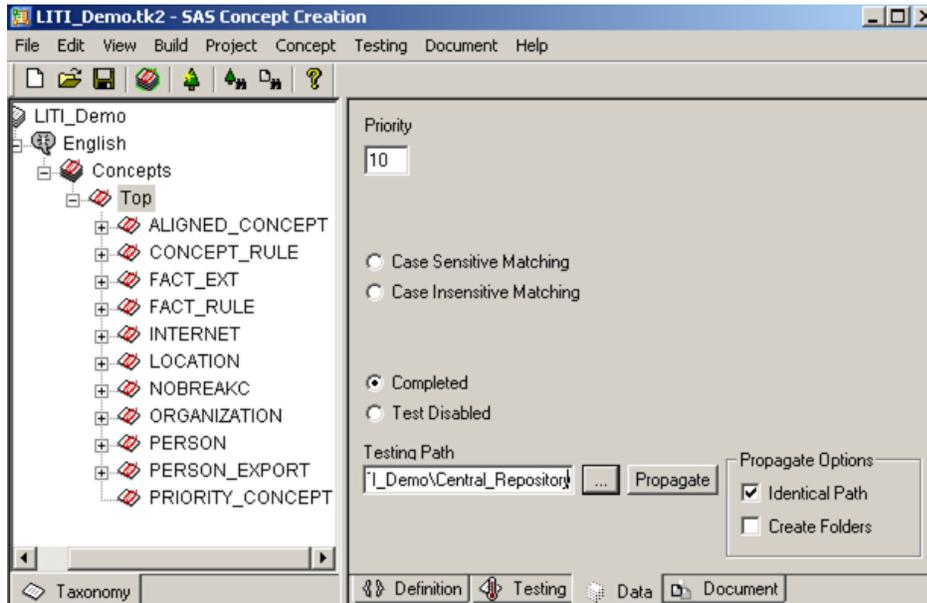
- This testing operation is typically the final testing stage and should replicate real-world results.
- These documents can be used to populate the testing taxonomy.
- This test operation can be a temporary substitute for a testing directory structure.

To create and set a path to the central repository, complete these steps:

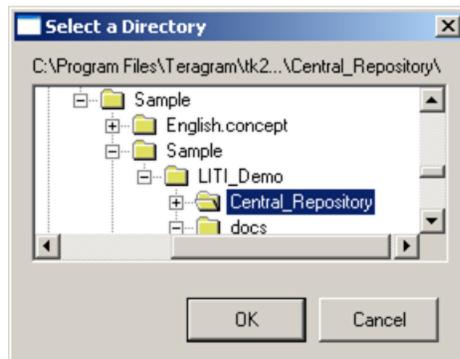
1. Create a single folder that is the central repository in the project directory on your hard drive. For example, create `Central_Repository`.



2. Select the `TOP` folder in the Taxonomy window.



3. Select **Identical Path** under the **Propagate Options** heading in the Data window.
4. Click  to the right of the **Testing Path** field and the Select a Directory window appears.

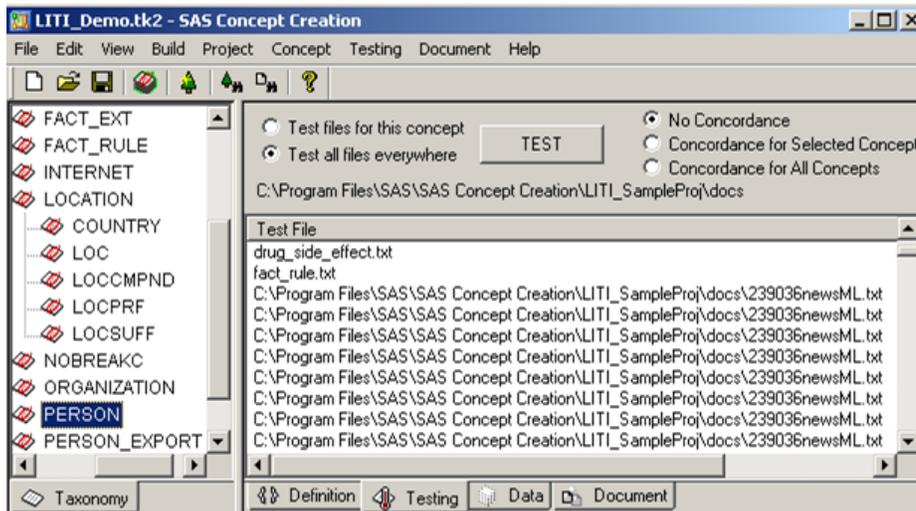


5. Select the central repository. For example, select `Central_Repository`.

6. Click **OK**.
7. Click **Propagate** in the **Data** tab. A SAS Concept Creation confirmation window appears.



8. Click **OK**. See the documents loaded into the **Testing** window.



9. (Optional) Click some of the concept nodes and you can see that each node displays the same path to the central repository in the **Data** window.

You can also choose to manually add additional documents to the central repository.

---

### 5.2.3 Manually Create a Testing Folder and Set a Path for a Newly Created Concept

If you add one or more concepts to the taxonomy, after you set up the testing directory, you can add a matching testing folder to the testing taxonomy. Manually set the path to this folder.

To add a test folder and set the path, complete these steps:

1. Access the testing directory. Create and name a new folder for the concept that you added to the taxonomy.
2. Enter the path to this folder into the **Testing Path** field of the Data window. Do not select either of the check boxes under **Propagate Options**.
3. Click **Propagate**. A SAS Concept Creation confirmation window appears.



4. Click **OK**.

## 5.3 Collecting Test Files

After you create repositories and set the paths to these directories, assemble different sets of testing documents. Choose texts that should be matched to the specific concepts that comprise your overall taxonomy structure.

The SAS Concept Creation testing process uses the testing taxonomy to determine the precision and recall of your concept extractor. Precision measures the relevancy of the matched documents, while recall measures whether all of the texts that should be returned are matched. For these reasons, each concept definition should be broad enough to include all of the texts that

---

you expect to match. These rules should also exclude any documents that do not belong to the selected concept.

Use the following steps to assemble the different types of texts required to test your taxonomy. In each case, choose documents of the types that are input to SAS Text Miner. For example, select `.html`, `.xml`, `.sgml`, `.pdf`, and `.txt` documents.

To assemble documents for individual concepts and for the central repository, complete these steps:

1. Select 10 or more documents that are matches for each concept in your taxonomy. These texts should have varying degrees of complexity levels for the definitions that you plan to match.
2. Copy and paste each group of documents into the testing folder named for the concept that they are expected to match. For more information, see Section 6.3 *Batch Testing* on page 146.
3. Collect a group of documents that include texts that are similar to the types of documents that are used when this application is applied in real time.
4. Copy and paste this group of texts into the central repository that you created. When you choose to use a central repository, you can see whether your documents match more than one concept and if so, why. For more information, see Section 6.5 *Test a Central Repository* on page 157.

---

## 5.4 Import Test Files

You can add additional testing files to the Testing window for a selected concept when you use the import test files operation. Use this operation with the central repository or any other testing folder.

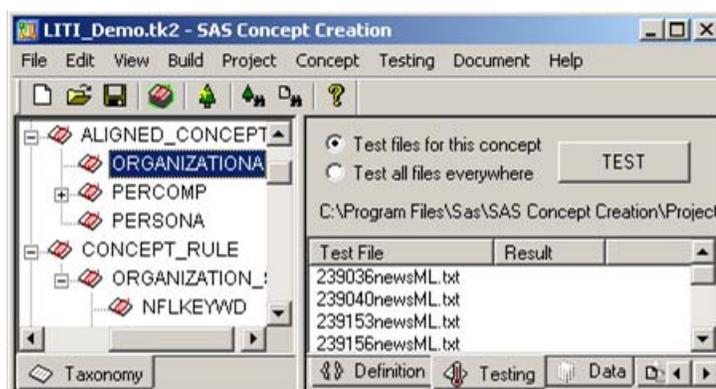
---

**Note:** Before you use the steps below, make sure that the Testing window is populated with some files.

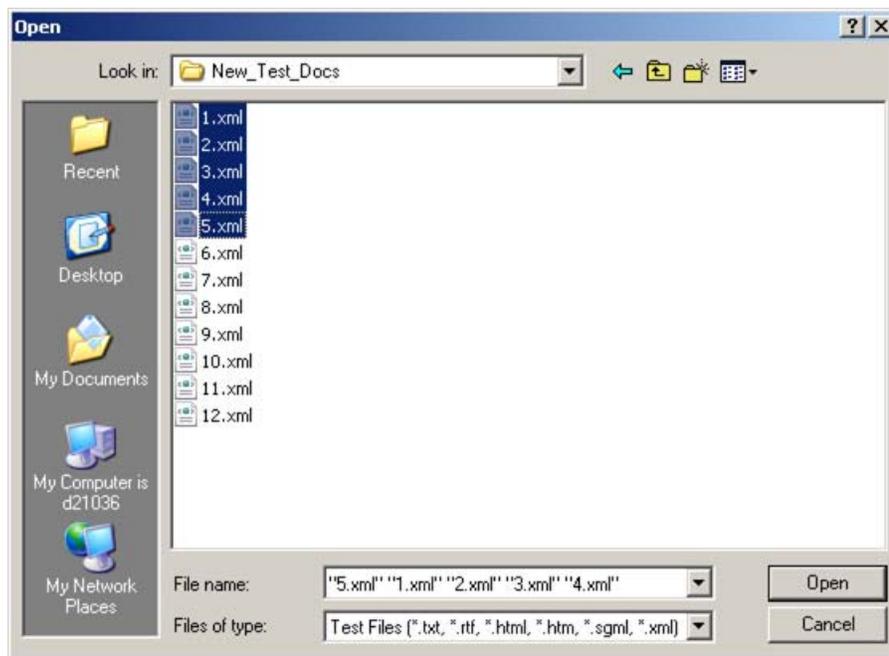
---

To import test files, complete these steps:

1. Select any concept and click the **Testing** tab. The **Test File** window displays the testing files that are found in the matched testing folder.

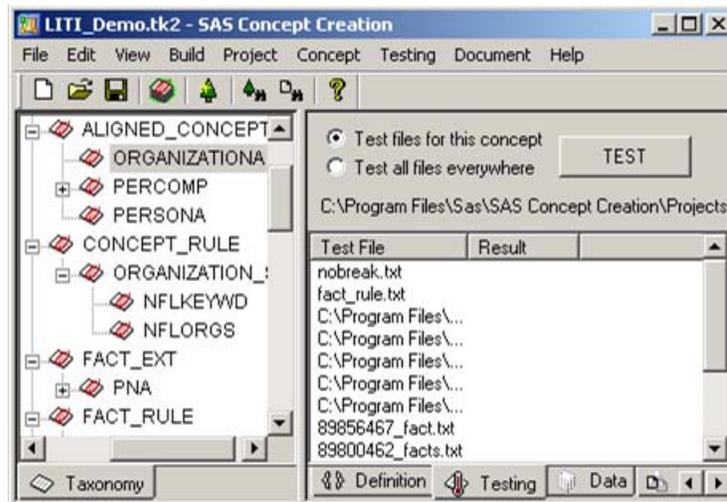


- 
2. Select **Testing --> Import Test Files** and the Open window appears.



3. Use Windows commands to select the test document, or documents, to add to the test operation.

4. Click **Open**. The selected test file, or files, is copied to the testing directory and listed in the **Testing** window.



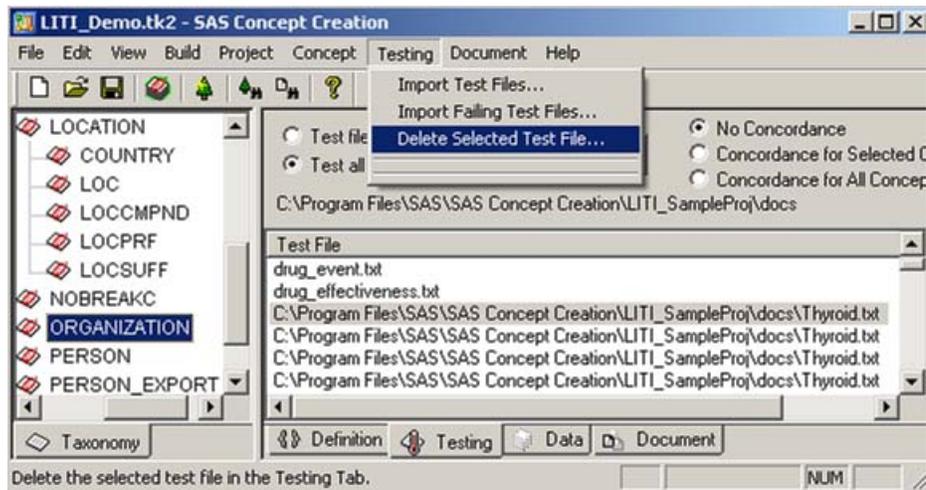
5. (Optional) Repeat Step 1 on page 139 through Step 4. above to add testing files to any other concepts.
6. Begin testing these files. For more information, see Section 6.3.2 *Option 1A: Batch Testing All of the Documents for One Concept* on page 147.

---

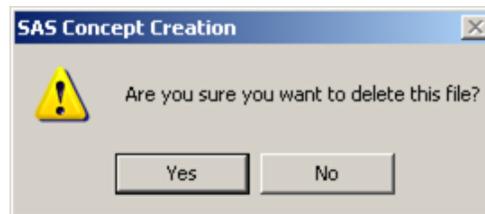
## 5.5 Delete Testing Files

To remove any of the testing files that you added to a testing folder, complete these steps:

1. Select a file in the **Testing** tab.



2. Select **Testing --> Delete Selected Test File**.
3. A SAS Concept Creation confirmation window appears.



4. Click **Yes**.

---

# Chapter: 6

## Testing the Concept Definitions

---

- *Overview of Testing*
- *Using the Testing Window*
- *Batch Testing*
- *Testing with the Document Window*
- *Test a Central Repository*
- *Comparing Test Results*
- *Import Failing Documents*
- *Testing with the Concordance*

### 6.1 Overview of Testing

Test the concept definitions that you develop in SAS Concept Creation for SAS Text Miner (SAS Concept Creation) before they are applied by SAS Text Miner. The testing process enables you to see how well your definitions perform and any necessary changes that are required before they are applied as custom entities in SAS Text Miner.

You can use different testing processes to examine the test results across the entire taxonomy, or choose to focus on matches within specific documents. Other testing processes include testing a single document against a selected concept or the entire taxonomy. You can also choose to create a folder of documents that should fail, but might not. For example, gardening documents should not include matches on the *Tournament of Roses*.

---

## 6.1.1 Windows

Use the following windows to test your concepts:

### Testing

Batch test the testing directory using the Testing window.

### Document

Select the Document window to test and see the testing results for a single document. You can test one document against a single concept or against all of the concepts in your project.

### Best Matches

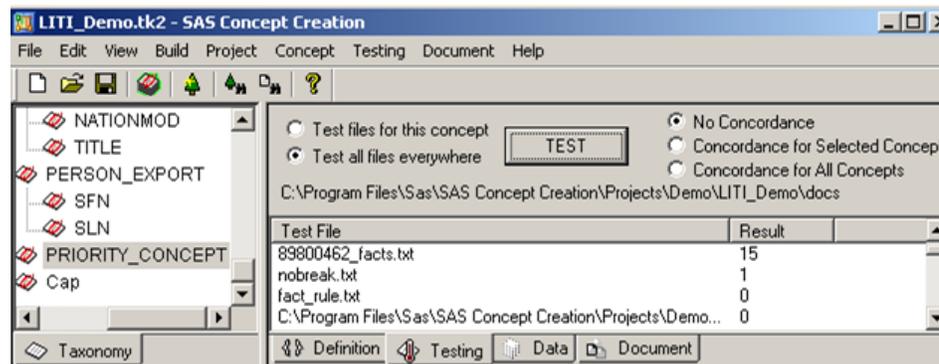
When you test against all of the concepts in your taxonomy, the Best Matches window appears.

## 6.2 Using the Testing Window

### 6.2.1 The Testing Window Messages

Before you use the Testing window, you should understand the types of information that appear. For information about the components of the Testing window, see Section 2.6.3 *The Testing Tab* on page 18.

*Display 6-1 Testing Window*



---

The following types of messages are displayed in the **Testing** tab:

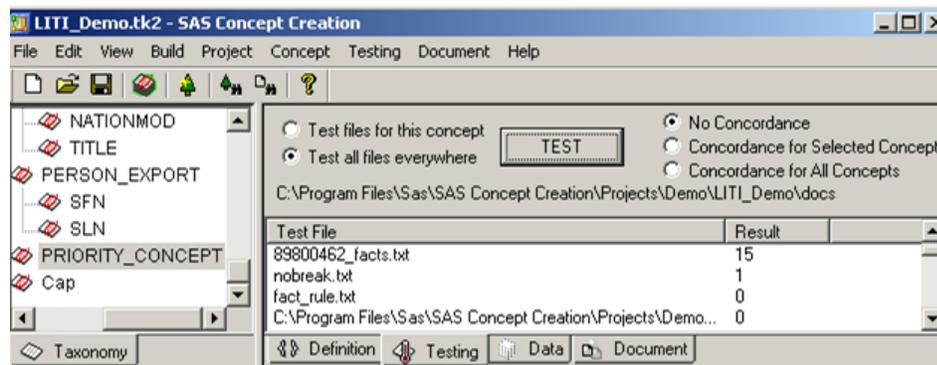
Path to the testing set of documents

This path appears below the **TEST** button and above the **Test File** heading. For example, you might see a path similar to this path:

```
C:\Program Files\SAS\SAS Concept  
Creation\Projects\Demo\LITI_Demo\docs
```

### Test File

See the list of test files that are tested in this window. The test files without a path belong to the testing folder that is matched to the selected concept. If a test file is followed by a path, it is an out-of-concept test file. These test files are imported using the **Test all files everywhere** or the **Testing --> Import Test Files** operation.



### Missing folders and files

#### No testing folder

If there is no testing folder that matches the selected concept in the testing taxonomy, a message such as `This directory does not exist` is displayed. Set the path to the testing directory using Section 5.2.1 *Create a Testing Directory While You Set Paths* on page 128.

#### Testing folder is empty

If the testing folder is empty, the message `No files found` appears. Copy test files into the testing directory.

### Result

The number of matching terms in the document appears.

---

## 6.3 Batch Testing

### 6.3.1 Overview of Batch Testing

A batch of testing documents is defined as the group of texts that you assemble to test. Before you begin to gather and test these documents, you should define at least some of the concepts in your taxonomy.

When you test multiple concepts using batches of testing documents, you gain information about the precision and recall of each definition. However, if testing documents that are not expected to match a concept do match, one of these rules might be too broad. If, on the other hand, the texts selected for the specified concept fail to match, the rule could be too narrow.

Batch testing, or testing one group of documents at a time, is only one of the testing operations available in SAS Concept Creation. Use a combination of these operations to develop a step-by-step, customized testing process that meets the specific requirements of your organization:

- Batch test your documents using the following operations in the Testing window:

**Test files for this concept**

Batch test all of the files that you selected for each concept against its definition. The test files that you assembled should pass the membership requirements for this concept. For more information, see Section 6.3.2 *Option 1A: Batch Testing All of the Documents for One Concept* on page 147.

**Test all files everywhere**

Use all of the documents in the testing directory. This means that you test all of the documents matched to each of the concepts in the taxonomy at one time, and against one concept. For more information, see Section 6.3.3 *Option 1B: Batch Testing the Testing Taxonomy or Out-of-Concept Files* on page 148.

- Use the Document window to see the matching results for one document. For more information, see Section 6.4 *Testing with the Document Window* on page 150.
- Test all of the documents in the central repository. This folder contains documents that should, and should not, match the selected concept. In

---

this case, you obtain test results that might be closer to the real project application. For more information, see Section 6.5 *Test a Central Repository* on page 157.

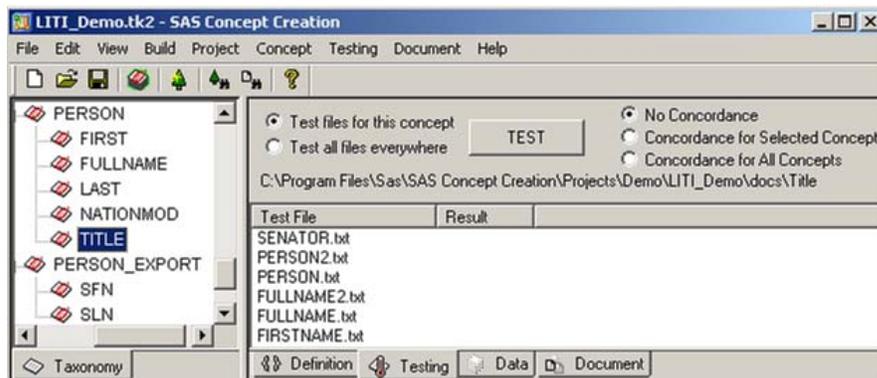
- Import failing test files at any time during the testing process. Failing test files are defined as documents that could pass, but should fail. For example, documents that mention *President George W. Bush* should not match definitions for concepts such as *Gardening bushes*. For more information, see Section 6.7 *Import Failing Documents* on page 159.

In summation, the batch testing operation provides an overview of the precision and recall of the concept definitions. This is true whether you test against a single taxonomy node or the entire taxonomy.

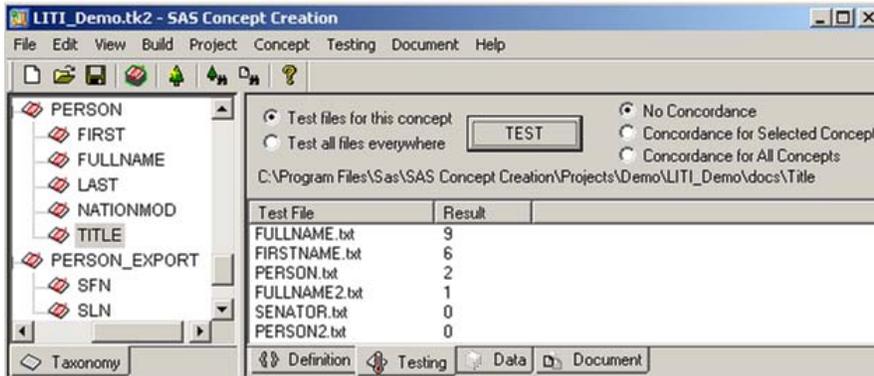
### 6.3.2 Option 1A: Batch Testing All of the Documents for One Concept

To batch test a testing set of documents against the concept that they are selected to match, complete these steps:

1. Create a testing taxonomy for your testing documents and set your testing paths. For more information, see Section 5.2.1 *Create a Testing Directory While You Set Paths* on page 128.
2. Select and assemble your testing documents. For more information, see Section 5.3 *Collecting Test Files* on page 137.
3. Select a concept to test in the Taxonomy window. For example, double-click on `TITLE`.



4. Click the **Testing** tab where the list of testing documents for this concept is displayed under the **Test File** heading. (In order to ensure the accuracy of your test file location, the path to the testing directory appears above the **Test File** heading.)
5. Select **Test files for this concept**.
6. Click **TEST**. The testing results appear in the Testing window.



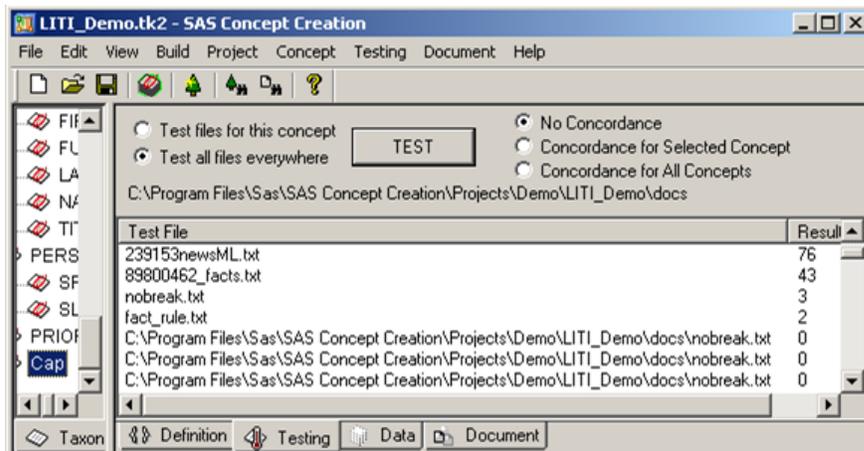
### 6.3.3 Option 1B: Batch Testing the Testing Taxonomy or Out-of-Concept Files

Batch test the entire testing taxonomy to see how test files selected for other concepts in the taxonomy perform. Analyze these test results to decide whether to make changes to your definitions.

To test all of the files in the testing directory, complete these steps:

1. Use Step 1 on page 147 through Step 4 on page 148.
2. Select **Test all files everywhere**.
3. Click **TEST**.

4. See the number of matching terms under the **Results** heading.



The testing files fall into one of two types:

#### In-concept files

These are the testing files that you assembled as optimal matches for the selected concept. When these names are displayed in the Testing window, no paths to these files are displayed. Instead the path to this testing folder is shown above the **Test File** heading and below the **TEST** button.

#### Out-of-concept files

Members of other testing folders are displayed with their full paths.

5. (Optional) To reverse the testing document ordering, click the **Test File** heading.
6. Compare the testing results for both types of files.

---

## 6.4 Testing with the Document Window

### 6.4.1 Overview of Document Window Operations

After you batch test a folder of testing documents against the concept that these texts were selected to match, test one document. This operation provides more detailed data by enabling you to see the matching terms for the selected concept within the document. In contrast, when you test all of your documents in the Testing window, you see a list of passing and failing texts.

You can also test this text against all of the concepts in the taxonomy. In either case, you see what terms matched in the Document window. Use the match highlighting to see what changes should be made to the definitions.

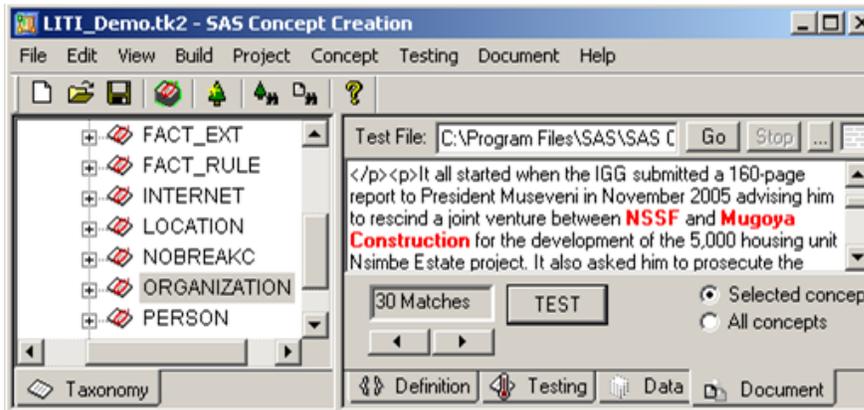
Test Web documents using the Document window as a browser. When you select this operation, you can remove the markup tags. Select **Document --> Remove Tags** to see the text without any markup language.

---

## 6.4.2 Test Using the Document Window

To test a document in the Document window, complete these steps:

1. Double-click on a document in the Testing window and this text appears in the Document window.



2. By default, you see the test results for the **Selected concept** displayed in the document. Use the matching terms, highlighted in red, to see the terms that made this document a passing text for the selected concept.
3. See the total number of matches displayed for this document. For example, see 20 Matches.
4. Click the  and  to jump through each of the matches in the window.

---

**Hint:** If you do not see the results that you expect, check your project settings. For more information, see Section 2.7.2 *The LITI Tab* on page 26.

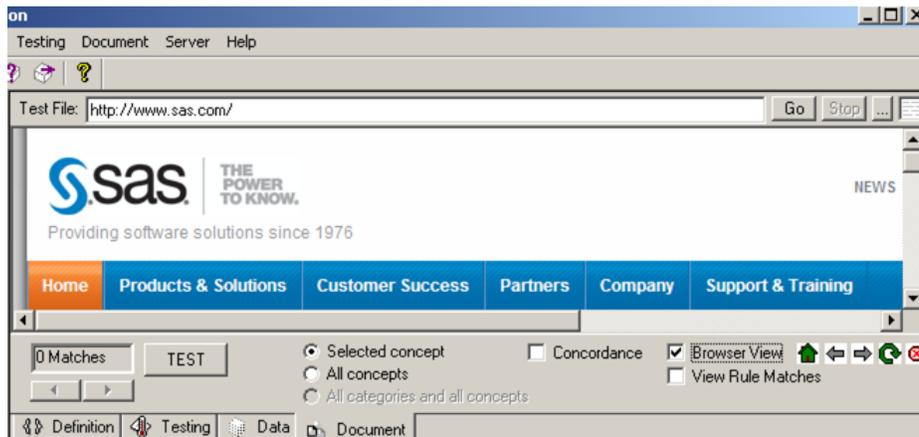
---

---

### 6.4.3 Testing a Web Page in the Document Window

Use the Document window to view Web pages. Also use this window to access operations that are specific to a Web browser such as viewing and testing Web pages, removing markup tags, and so on. Select **Browser View** to access these operations.

*Display 6-2 Web Page in Browser View*



---

**Note:** Web pages are tested in their source format.

---

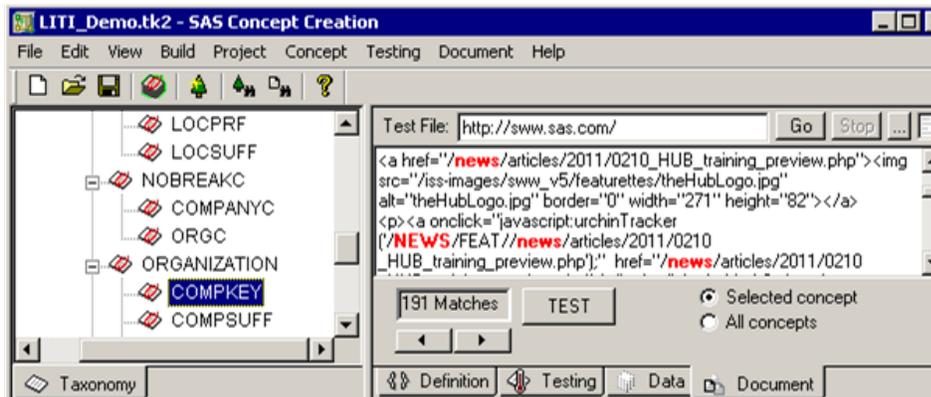
The browser operations are described in Table 2-3 on page 24.

To test a Web page as a text document, complete these steps:

1. Select a concept in the Taxonomy window. For example, select **COMPKEY**.



2. Click the **Document** tab.
3. Select **Browser View**.
4. Enter the URL of the Web page that you want to test into the **Test File** field.
5. Click **TEST**. The results of the testing operation appear in the source document.



6. (Optional) Select **Document --> Remove Tags** to delete the markup tags such as `</p>` in the source document. (If you perform this operation, click **TEST** to see the tags reinstated.)

---

**Hint:** Before you remove any tags, check your **Project Settings - Misc** tab and any of your rules that specify a `PARA` tag.

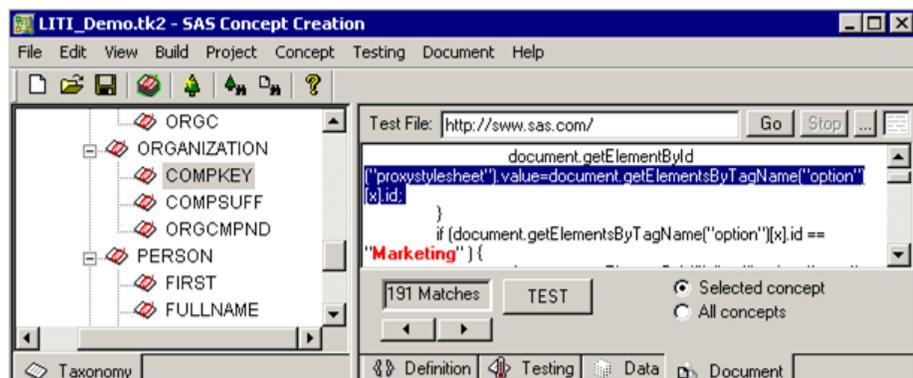
---

## 6.4.4 Using Windows Commands

You can use the delete and replace text commands in order to change the text in your documents. Use these operations to see how changes affect matching in the input documents.

To remove text from your testing document, complete these steps:

1. Double-click on a document in the **Testing** tab and the contents of the text appear in the **Document** tab.



2. Highlight the text that you want to delete and press the Delete key on your keyboard.
3. (Optional) Enter any words that you want to add to the document.
4. Click **TEST** to see whether you obtain the results that you require.

---

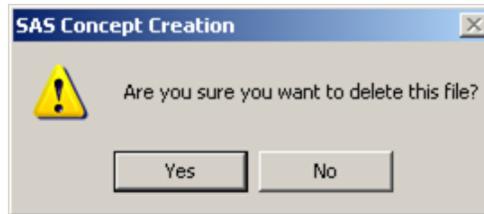
**Note:** The Document window is not a text editor. For this reason, any changes that you make are not permanent. The original document remains intact both

---

as it is tested in the Testing window and in the testing folder.

---

5. When you leave the Document window and try to test another document, a SAS Concept Creation confirmation window appears.



6. Click **Yes**, unless you want to continue to test document in the Document window.

### 6.4.5 Copy and Paste a Test File

You can copy and paste a test file directly into the Document window. Use this operation to test a text without including it in the test file folder.

To copy and paste a test file, complete these steps:

1. Access the Document window.
2. Access another document that you want to test in *Notepad*.
3. Highlight the text that you want to test.
4. Copy this text, or the whole document. Use Ctrl V to paste the text into the Document window.
5. Click **TEST** to test the document.

### 6.4.6 Using Clear Test Document

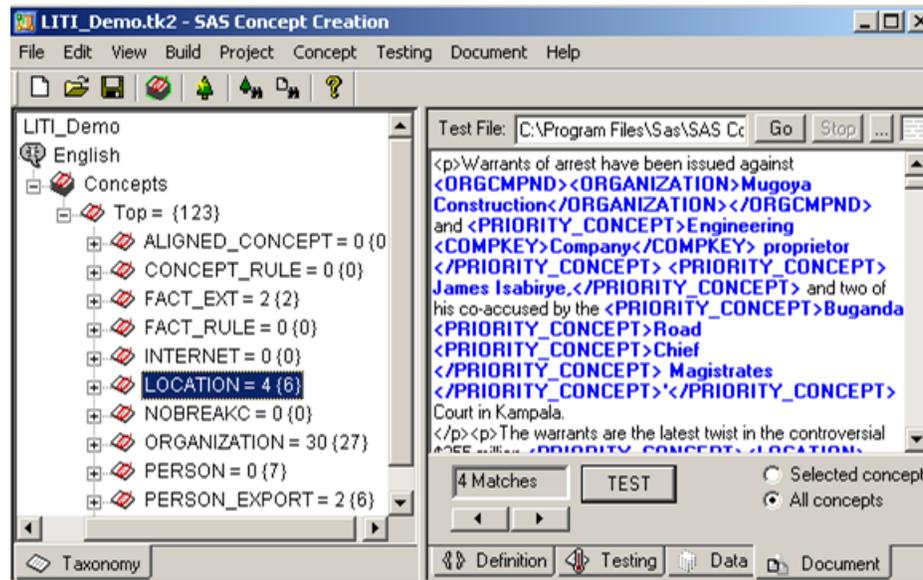
When you select the **Document --> Clear Test Document** operation, the document that currently appears in the Document window is removed from the Document window. However, this text is not deleted from the list that appears in the Testing window or from the testing folder.

---

## 6.4.7 Refreshing the Taxonomy Tree

Refresh your taxonomy tree when you want to retest your document by deleting all of the numbers that indicate the matches that appear in the Taxonomy window. These messages appear after you test a text using the **All concepts** radio buttons in the **Document** tab.

*Display 6-3 An Example of Match Counts*



To delete the `PASS` and `FAIL` messages in the Taxonomy window, click the **Refresh Tree** button, or access a new document in the Document window.

## 6.4.8 Changing the Font Size of a Tested Document

You can choose to increase or decrease the size of the text that is displayed in the Document window. These operations can make it easier to see the matching terms within their context.

To increase the font size, select **Testing --> Increase Font Size**.

To decrease the font size, select **Testing --> Decrease Font Size**.

---

**Note:** You can decrease the size of the letters in the document only after you have increased their size.

---

### 6.4.9 Removing Markup Tags

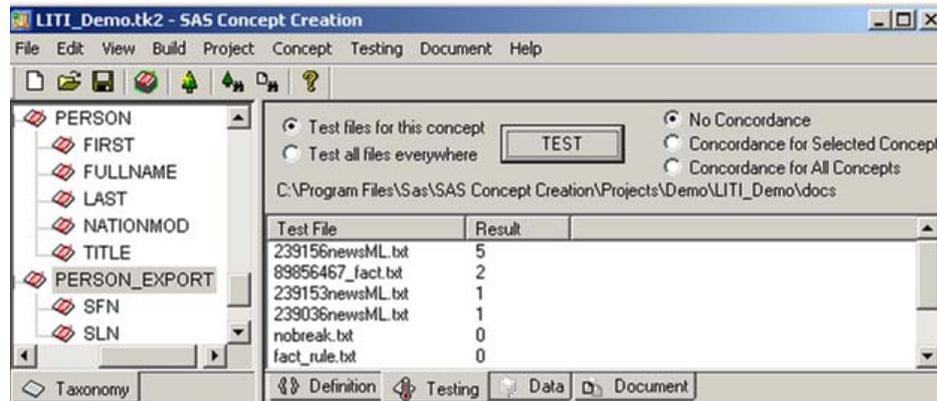
To see an HTML, or an XML, document as a text without any markup tags, select **Document --> Remove Tags**. The testing document in the Document window is displayed as a text document without any markup language.

## 6.5 Test a Central Repository

The central repository is a collection of testing documents that are not selected to match any specific concepts.

To test the central repository, complete these steps:

1. Use the steps in Section 5.2.2 *Create and Set a Path to the Central Repository* on page 134.
2. Select a concept in the Taxonomy window and click the **Testing** tab.



3. Click **TEST** to see the test results.

- 
4. See the results and compare them to those that you obtained testing the selected documents for this concept.

---

**Hints:** The testing path is the same for each of the concepts in the taxonomy.  
The list of testing documents is also identical.

---

## 6.6 Comparing Test Results

The testing results that are displayed in the Testing window for both *in-concept* and *out-of-concept* files enable you to compare the test results. These results provide a more comprehensive view of the appropriateness of your definitions. For example, if one of the passing documents for the FIRST concept matched the FULLNAME concept, you should determine why this unexpected behavior occurred. Analyze the FIRST definition for the purposes of understanding why this document matched. Also examine the LAST rule and the matched document. One, or both, of these definitions might be too broad.

If you double-click the tested document, the text appears in the Document window. Examine the matched terms in this window to gain a better understanding of why this document matched the FIRST rule. For more information, see Section 6.4 *Testing with the Document Window* on page 150.

Conduct additional testing to evaluate the performance of other documents. Further testing could identify whether you should take one or more of the following actions:

- Narrow a concept rule. For example, remove the term *Basketball* from the Sports concept rule.
- Broaden the concept rule. For example, add one or more of the terms that are used to define the PERSON concept rule to the ORGANIZATION rule.
- Eliminate one, or more, of these concepts from your taxonomy.
- Add additional concepts to your taxonomy structure. For example, add a child node below the ORGANIZATION concept that is *PUBLIC*.

---

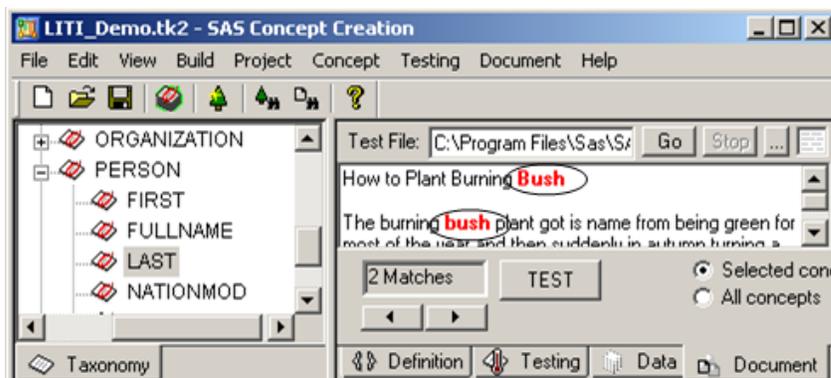
**Note:** When you perform any of these operations, test your results after each step in the process.

---

## 6.7 Import Failing Documents

During testing, you might discover that certain test documents should not be matched to a specific concept. For example, landscaping texts that contain the word *bush* should not match the LAST concept that contains a rule for president names.

Figure 6-1 Failing Document Example

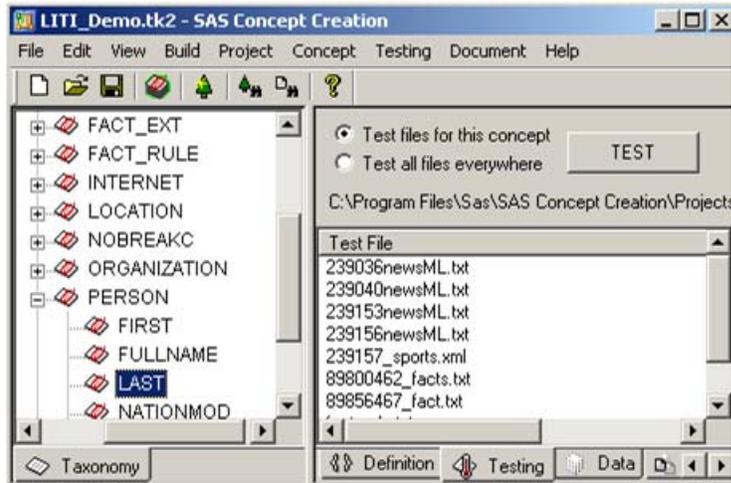


In the example provided above, the passing document entitled *How to Plant Burning Bush*, contains the word *bush* in the context of a plant. This is an example of a document that you do *not* want to pass the test for the LAST concept where one of the classifier rules specifies the word *bush*.

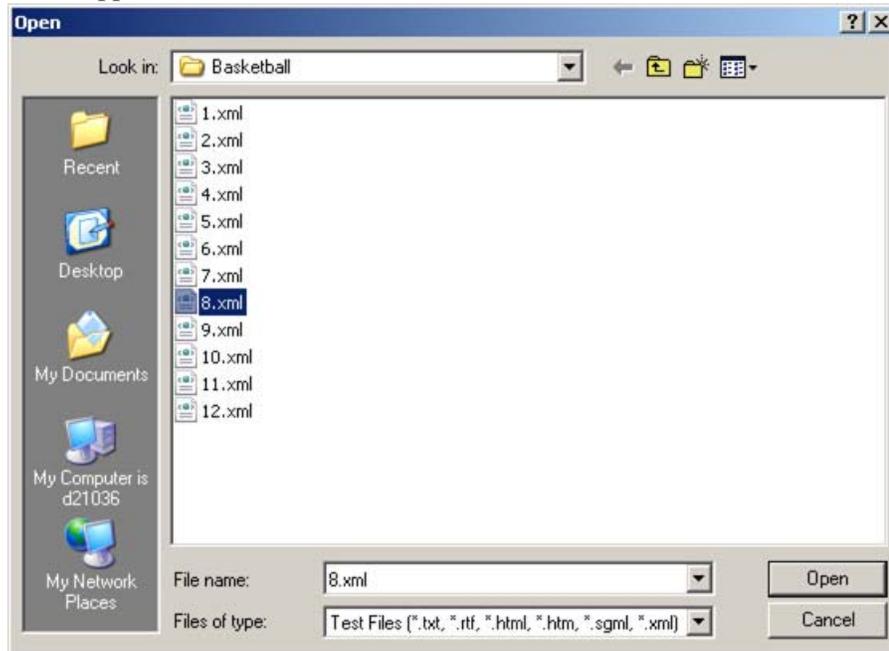
As you test and define concept definitions, copy documents that should fail, but are not, into a Fail directory. You can test this directory as a final step in the testing process to confirm the accuracy of your definitions.

To test documents in the Fail directory, complete these steps:

1. Click the **Testing** tab.



2. Select **Testing --> Import Failing Test Files**. The Open window appears.



- 
3. Select a file. For example, choose 8.xml.
  4. Click **Open**. The failing testing document appears in the Testing window preceded by its path.
  5. Click **TEST** to see whether this file fails, or whether you need to make further rule adjustments. See the example provided in Figure 6-1 on page 159.

## 6.8 Testing with the Concordance

### 6.8.1 An Overview of the Concordance

The concordance feature enables you to see a list of the matched terms highlighted in red, in one or all, of the input documents. You can choose to use the concordance operations that are available in both the Testing and the Document windows.

Use the concordance operations that are available in the Testing window to see the concordance matches that are returned for all of the documents listed in this window. Use the concordance operations that are available in the Document window to see matches displayed within the text of the document.

### 6.8.2 The Concordance for the Testing Window

Select one of the following operations in the Testing window in order to see the different types of concordance matches:

**Concordance for Selected concept**

see all of the matches for the selected concept in the Taxonomy window.

**Concordance for All concepts**

see all of the matches for all of the concepts in this project.

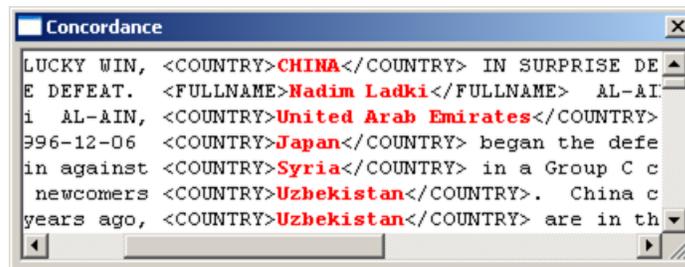
The results are displayed according to the selections that you specify. These selections include the operations specified in the **Project Settings - Concepts** window. For more information, see Section 6.8.3.B *Determine How the Concordance Is Displayed* on page 163. Also see Section 2.8.4 *The Concordance Windows That Are Available through the Testing Tab* on page 35.

See the following examples:

*Display 6-2 Concordance for Selected Concept*



*Display 6-3 Concordance for All Concepts*



## 6.8.3 The Concordance for the Document Window

### 6.8.3.A An Overview of the Concordance for the Document Window

Select the **Concordance** check box and make one of the following selections in the Document window in order to see the different types of concordance matches:

#### **Selected concept**

see all of the matches for the selected concept.

#### **All concepts**

see all of the matches for all of the concepts in this project.

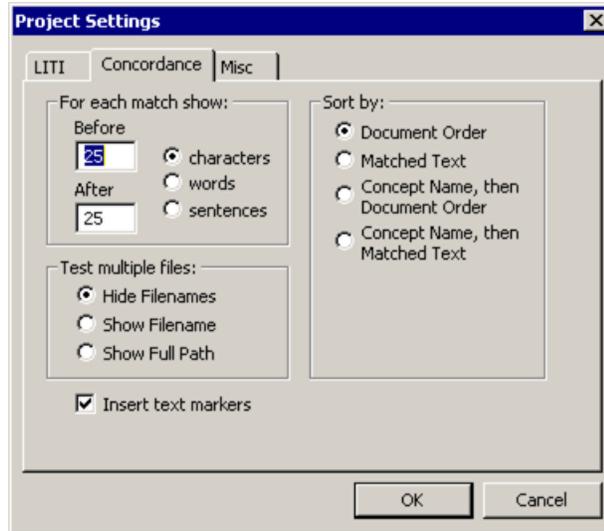
---

The results are displayed according to the selections that you specify. These selections include the operations specified in the **Project Settings - Concepts** window.

### 6.8.3.B Determine How the Concordance Is Displayed

To set up the display for the concordance, complete the following steps:

1. Select **Project --> Settings**.
2. Click the **Concordance** tab.



3. Under **For each match show**, select your settings.
4. Select a **Sort by** operation.
5. Determine how to **Test multiple files**:
6. Select **Insert text markers** to display text markers in the concordance view of the **Document** tab when you test a single file against multiple concepts. The match text fields display the concept that is the best match for the matched term that is returned. One example of these tags is `<CONCEPT1> . . . </CONCEPT1>`.

---

**Note:** For more information, see Section 2.7.4 *The Concordance Tab* on page 29

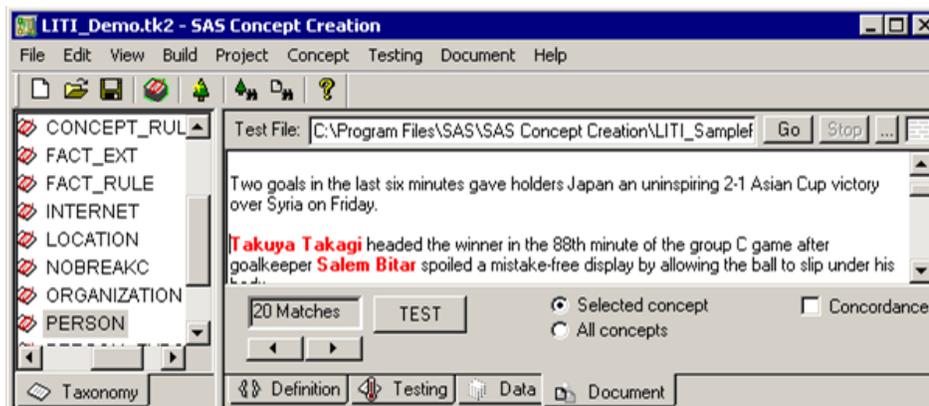
---

### 6.8.3.C See the Concordance Terms for a Selected Concept

Use the concordance to see a list of the terms in the input document that match only the selected concept.

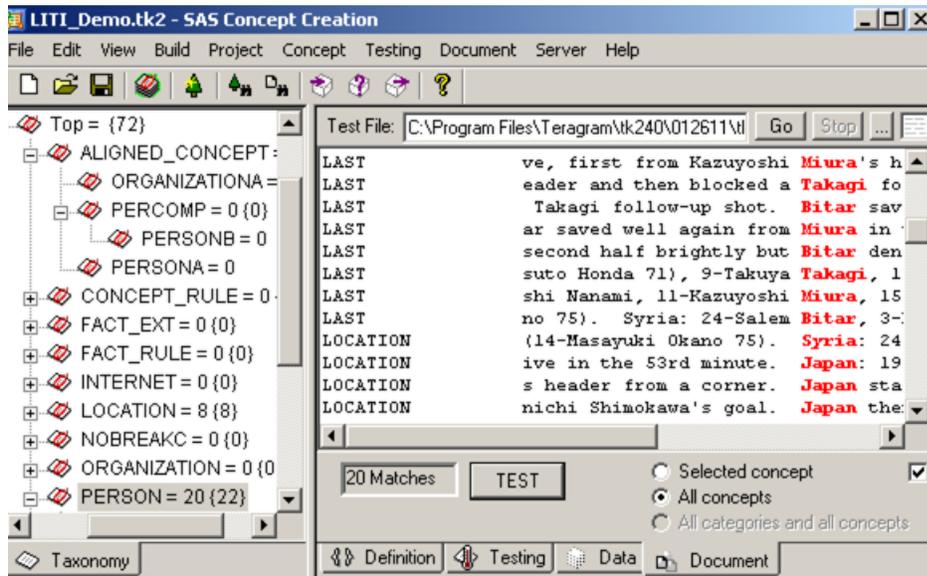
To see a list of matching terms for a selected concept, complete these steps:

1. Test the testing documents for a selected concept in the **Testing** tab.
2. Double-click a tested document and it appears in the **Document** tab.



3. By default, **Selected Concept** is selected. If not, choose **Selected Concept**.
4. Select **Concordance**.
5. Click **TEST**.

- See the matching terms highlighted in red. The highlighted terms, and whether a matched concept is displayed, depend on the selection that you choose in the Project Settings - Concordance window.



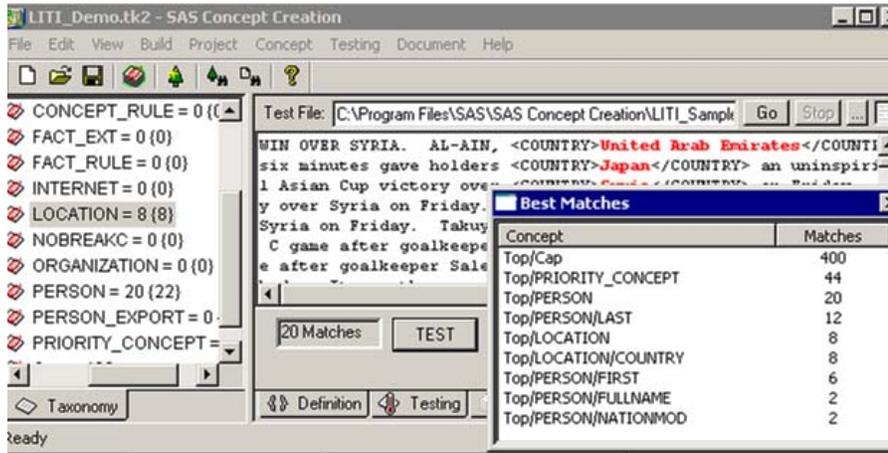
### 6.8.3.D Use the Best Matches Window for All Concepts

When you choose to see all of the matching terms in an input document for your taxonomy, you can also see these results in the Best Matches window.

To see a list of matching terms for all of the concepts in the taxonomy, complete these steps:

- Access a test document in the **Document** tab. For more information, see Section 6.4.2 *Test Using the Document Window* on page 151.
- Select **Concordance**.
- Click **TEST**.
- See the matching terms highlighted in red and preceded by the name of the matched concept.

5. See the matches for all concepts in the taxonomy in the Best Matches window.



The Best Matches window ranks the matching concepts according to the total number of matches that occur for this all concepts. These totals are listed from the highest to the lowest numbers under the **Matches** heading. For more information, see Section 2.8.11 *The Best Matches Window* on page 42.

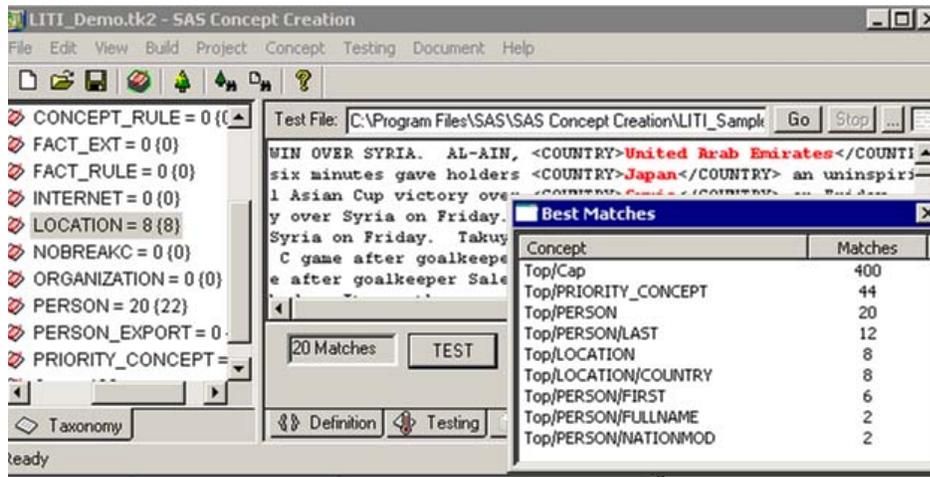
### 6.8.3.E See the Concordance Terms for All Concepts

You can choose to see all of the matching terms in an input document for all of your concepts. When you select this operation, you can also see the results in the Best Matches window.

To see a list of matching terms for all of the taxonomy nodes, complete these steps:

1. Access a test document in the **Document** tab. For more information, see Section 6.4.2 *Test Using the Document Window* on page 151.
2. Select **All concepts**.
3. Select **Concordance**.
4. Click **TEST**.

- See the results in both the concordance and Best Matches window. For more information, see Section 2.8.11 *The Best Matches Window* on page 42.





---

# Appendixes

---

- Appendix A: *Regex Syntax and Part-of-Speech on page 171*
- Appendix B: *Glossary on page 181*
- Appendix C: *Recommended Reading on page 185*



---

# Appendix: A

## Regex Syntax and Part-of-Speech

---

- *Regular Expressions*
- *Part-of-Speech Table*

### A.1 Regular Expressions

#### A.1.1 Rules and Restrictions

The following rules and restrictions apply to regular expressions:

- Any single character **a** (ASCII 1 through 255, subject to escaping restrictions in 14 below) is a regular expression, and it matches precisely that character.
- A character class is a regular expression. One or more characters inside square brackets (**[ ]**), match any of the characters specified inside of the square brackets. For example, **[abc]** matches **abc**. A range inside a character class such as **a-z** matches any ASCII character whose value is between **a** through **z**, inclusive. Any character, including special characters, can appear in a character class. However, **\** (backslash), **-** (hyphen), **[** and **]** (open and closed brackets) are preceded by a backslash. If you want to return a literal match on these characters, see Section A.1 *Regular Expressions* on page 509.
- A negated character class is a regular expression. One or more characters are inside square brackets, with **^** (caret) being the first character to indicate negation. For example, **[^abc]** matches any character except **a**, **b**, or **c**. (If you want to return a literal match on a caret, precede the caret with a backslash.)

Also see the table below for more information about the rules and restrictions for regular expressions.

Table A-1: More Rules and Restrictions

If Statement	Explanation
If <b>a</b> and <b>b</b> are regular expressions	then so is <b>ab</b> that matches whatever <b>a</b> matches followed by whatever <b>b</b> matches (concatenation)
	then so is <b>a b</b> that matches either whatever <b>a</b> matches or whatever <b>b</b> matches
If <b>a</b> is a regular expression	then so is <b>(?:a)</b> that simply serves as a grouping mechanism without remembering what it was grouping. For example <b>(?:ababb) b</b> matches either <b>abaab</b> or <b>b</b> . This would be difficult to express without the grouping mechanism.
	then so is <b>a*</b> that matches 0 or more occurrences of whatever <b>a</b> matches
	then so is <b>a+</b> that matches 1 or more occurrences of whatever <b>a</b> matches
	then so is <b>a?</b> that matches 0 or 1 occurrences of whatever <b>a</b> matches
	then so is <b>a{n,m}</b> that matches at least <b>n</b> but no more than <b>m</b> concatenated occurrences of whatever <b>a</b> matches
	then so is <b>a{n,}</b> that matches at least <b>n</b> concatenated occurrences of whatever <b>a</b> matches
	then so is <b>a(n)</b> that matches exactly <b>n</b> concatenated occurrences of whatever <b>a</b> matches

---

## A.1.2 Special Characters

The table below lists, and gives extended meaning to, special characters that are used with regular expressions.

Table A-2: Special Characters in Regular Expressions

Character	Meaning
\a	Alarm (beep)
\n	Newline
\r	Carriage return
\t	Tab
\f	Form feed
\e	Escape
\d	Digit (same as <b>[0-9]</b> )
\D	Not a digit (same as <b>[^0-9]</b> )
\w	Word character (same as <b>[a-zA-Z_0-9]</b> )
\W	Non-word character (same as <b>[^a-zA-Z_0-9]</b> )
\s	Whitespace character (same as <b>[\t\n\r\f]</b> )
\S	Non-whitespace character (same as <b>[^\t\n\r\f]</b> )
.	Wildcard (matches any character)
\xh	Hexadecimal number, where <b>h</b> is a hexadecimal character
\xhh	Hexadecimal number, where <b>h</b> is a hexadecimal character
\oo	Octal number, where <b>o</b> is an octal digit
\ooo	Octal number, where <b>o</b> is an octal digit

---

## A.1.3 Special Cases

There are several special cases for regular expressions. These cases include:

[,],(,),?,\*,+,.,-,\\,|

for metacharacters such as these to have literal meaning, these metacharacters need to be escaped with a backslash (\). If inside a character class, however, only those metacharacters that are explicitly mentioned need escaping.

No support is provided for the following:

backward references

O as a remembering grouping mechanism.

^ as the beginning-of-line zero-width assertion

\$ as the end-of-line zero-width assertion

---

**Note:** Unlike Perl regular expressions, the ^ and \$ markers are implicitly assumed.

---

## A.2 Part-of-Speech Table

The table below provides a list of the part-of-speech tags that you can use to write rules. Also see the descriptions and examples included in this table. For more information about how these parts of speech are used to write rules, see Section 4.4.14 *The Part-of-Speech Tags* on page 74. Also see the language book for each language that you purchased.

---

**Note:** Use the part-of-speech tags that are listed here. Do not use the part-of-speech tags that are used with SAS Text Miner. The part-of-speech tags that are specified

in your definitions are mapped to those in SAS Text Miner at the time of application.

Table A-3: Part-of-Speech Morphological Features

Code	Part-of-Speech	Example
A	adjective	The sky is <i>azure</i> .
ABBREV	abbreviation	etc.
Acomp	comparative adjective	The green bag is <i>heavier</i> than the red one.
Adv	adverb	He is <i>easily</i> the best candidate.
Asup	superlative adjective	He cooked the <i>best</i> dish.
C	conjunction	Say nothing of former informers <i>and</i> spies.
date	valid date formats YYYY-MM-DD YYYYMMDD YY-MM-DD YYMMDD YYYY-MM YYYYMMs YY-MM Standard US Date Formats MM-DD-YYYY MM/DD/YYYY MM-DD-YY MM/DD/YY	04JAN2001 04jan2001
Det	determinant	Nothing can be further from <i>the</i> truth.
digit	numeric symbols, including floating point decimals	5, 2.14, or 5,254

Table A-3: Part-of-Speech Morphological Features (Continued)

Code	Part-of-Speech	Example
F	French word	We went to see the <i>chateaux</i> .
inc	unknown word to the part-of-speech tagger	
Int	interjection	Yum!
Md	modal verb	This <i>might</i> be the best idea.
Mdn 't	modal verb negated	I <i>won't</i> elaborate on this any further.
N	noun	The <i>e-mail</i> went to the spam folder.
Npl	plural noun	The <i>geese</i> are leaving for the South.
Num	number	She just turned <i>seventeen</i> years old.
PN	proper noun	We are going to <i>England</i> for vacation.
PossDet	possessive determinat	It is <i>her</i> choice.
PossPro	possessive pronoun	The choice is <i>hers</i> alone.
PreDet	<i>pre</i> determinat	<i>All</i> the king's soldiers could not put him together again.
Prefix	prefix	The <i>multi</i> -millionaire Soros is going to help us out.
Prep	preposition	Let's go <i>to</i> grandma's house.
Pro	pronoun	Give me one of <i>each</i> .
ProMD	pronoun contracted with modal	If it <i>weren't</i> for him, we'd still be here.
ProV	pronoun contracted with a verb	we're
Ptl	particle	I would go <i>across</i> if I could.

Table A-3: Part-of-Speech Morphological Features (Continued)

Code	Part-of-Speech	Example
RelPro	relative pronoun	I want the coin <i>that</i> represents King Kong.
sep	separator character	;,;,.,.
time	time formats 23:59:59 235959 23:59:59.9942 235959.9942 23:59:59Z 23:59:59.9942Z 235959.9942Z 23:59:59+HH:MM 23:59:59-HH:MM 235959+HHMM 23:59:59.9942Z 235959.9942Z	12:56:32

Table A-3: Part-of-Speech Morphological Features (Continued)

Code	Part-of-Speech	Example
time (continued)	Standard US and British Time Formats 10:15AM 10:15A.M. 10:15am 10.15a.m. 10AM 10A.M. 10am 10a.m. 10:15PM 10:15P.M. 10:15pm 10.15p.m. 10PM 10P.M. 10pm 10p.m.	9:00PM
url	urls	www.sas.com/success/
v	verb	You should <i>verbalize</i> your wishes.
V3sg	verb, 3 <sup>rd</sup> person singular	The boy <i>amuses</i> himself throwing rocks.
V3sgn't	verb, 3 <sup>rd</sup> person singular negated	This <i>isn't</i> funny.
Ving	present participle	Why is the hen <i>crossing</i> the street?
Vn't	negated verb	"it <i>don't</i> mean a thing..."

Table A-3: Part-of-Speech Morphological Features (Continued)

Code	Part-of-Speech	Example
Vpp	past participle	Those tapes were <i>released</i> .
Vpt	verb, past tense	The president <i>hated</i> broccoli.
Vptn't	verb, past tense negated	If it <i>weren't</i> for him, we'd still be here.
WAdv	w adverb	<i>Why</i> do you say that?
WDet	w determinant	<i>What</i> is he saying?
WPossPro	w possessive pronoun	<i>Whose</i> hat is this?
WPro	w pronoun	<i>Whom</i> did you meet?



---

# Appendix: B

## Glossary

---

**\_c**

specifies the context for the matches.

**\_cap**

specifies that a word beginning with an uppercase letter is a match.

**argument**

is defined by two or more concepts that are related to each other. When these matches are identified, arguments are returned. See *fact*.

**canonical form**

specifies the full name, or form, of the term. For example, SAS Institute Inc. is the canonical form of SAS.

**CLASSIFIER**

specifies the terms to be matched.

**CONCEPT**

locate entities, or ideas, in input documents.

**concordance**

displays a list of the matching terms located in a document with the text surrounding them. Specify the number of characters or words that are returned when a match on a concept occurs.

**coreference**

refers to pronoun resolution. A pronoun is matched to the antecedent that it refers to. Coreference is also known as *anaphora resolution*.

**definition**

defines a concept, whether it consists of one or more rules. *Definition* is used interchangeably with the word *rule*. See *rule*.

---

**event**

is used interchangeably with *fact*. See *fact*.

**Fact**

refers to two or more concepts or tokens that are specified in one `SEQUENCE` or `PREDICATE_RULE` definition. See *SEQUENCE* and *PREDICATE\_RULE* below.

**precision**

is a measurement of the relevancy of the matched documents. In other words, the concept definition excludes possible matches that do not reflect the subject matter of the concept. For example, texts referring to *rock collections* are not matched for the concept *Rock and Roll*.

**PREDICATE\_RULE**

returns matches when an operator is specified with arguments. Unlike the `SEQUENCE` definition, the matches do not need to occur in the order specified by the definition.

**SEQUENCE**

returns facts when matches occur within the specified context.

**priority**

ranks concepts. By default, priority is set to 10 in the Data window.

**recall**

a measurement of how well the definition matches all of the relevant texts.

**referring term**

a term that refers to a canonical form.

**REGEX**

specifies regular expression syntax.

**rule**

defines the concept. There can be many rules for each concept definition. This term is used interchangeably with *definition*, but properly speaking, one definition can contain many rules. See *definition*.

**SEQUENCE**

returns Facts when matches occur within the specified ordering.

---

**string**

refers to a group of words or characters that you specify for a rule.

**token**

a synonym for a word. *Token* is not a synonym for the word *string* that can refer to several words or characters. *Token* refers to one word, only.



---

# Appendix: C

## Recommended Reading

---

The following books are recommended as companion guides:

- *SAS Concept Creation for SAS Text Miner: Installation Guide*: Install SAS Contextual Extraction Studio
- *Getting Started with SAS Text Miner*: Learn how to use SAS Text Miner.
- *Installation and Upgrade Instructions for SAS Text Miner*: Install SAS Text Miner before you create a project using SAS Concept Creation for SAS Text Miner.
- Use the language books for each language purchased to see the comprehensive list of part-of-speech tags that are available.
- SAS offers instructor-led training and self-paced e-learning courses to help you get started with the SAS add-in, learn how the SAS add-in works with the other products in the SAS Enterprise Intelligence Platform, and learn how to run stored processes in the SAS add-in. For more information about the courses available, see [support.sas.com/training](http://support.sas.com/training).

For a complete list of SAS publications, see the current SAS Publishing Catalog. To order the most current publications or to receive a free copy of the catalog, contact a SAS representative at

SAS Publishing Sales  
SAS Campus Drive  
Cary, NC 27513  
Telephone: (800) 727-3228\*  
Fax: (919) 677-8166  
E-mail: [sasbook@sas.com](mailto:sasbook@sas.com)  
Web address: [support.sas.com/pubs](http://support.sas.com/pubs)

\* For other SAS Institute business, call (919) 677-8000.

Customers outside the United States should contact their local SAS office.



---

# Index

---

%	usage .....	75
+	usage .....	75
>	usage .....	71, 87, 115
_c	context operator .....	115
	usage .....	70, 86
_cap	defined .....	68
	usage .....	71
_coref	classifier rule .....	114
_F	usage .....	116
_P	usage .....	117
_ref	export symbol .....	119
	new concept .....	118
	usage .....	114, 115
_w	usage .....	70
{}	usage .....	108

## A

Abort Compiling Concepts	
Build menu .....	9
Add Concept	
Concept menu .....	10
usage .....	33, 48
Add Language	
Project menu .....	9

---

ALIGNED	
defined	79
usage	80
All Concepts	
define	21
defined	25
All matches	
Data window	105
usage	60, 110
AND	
defined	79
usage	80
argument	
defined	107
fact	113

## B

Back button	
Browser	12
Web view	25
batch testing	
benefits	147
defined	146
Best	
Data window	106
usage	60
Best Matches window	
Document window	144
usage	42
Browser	
Back button	12
Forward button	12
Home button	12
Refresh button	12
Stop button	12
Browser View	
defined	25
Testing menu	12
usage	23, 152

---

Build menu	
Abort Compiling Concepts .....	9
Compile Concepts .....	9

## C

C_CONCEPT	
_ref operator .....	114
defined .....	68
spaces .....	73
canonical form	
coreference .....	114
Case Insensitive Matching field	
defined .....	20
Case Sensitive Matching field	
defined .....	20
case-insensitive	
matching .....	69
case-sensitive	
matching .....	67
central repository	
defined .....	146
testing .....	146
CLASSIFIER	
coref .....	118
defined .....	68, 83
classifier rule	
_coref .....	114
Clear Test Document	
Testing menu .....	11
colons	
usage .....	73
commas	
usage .....	72
Compatibility Date	
Misc window .....	28
Compile Concepts	
Build menu .....	9
Compile Concepts tab	
usage .....	39

---

Completed field	
defined .....	20
CONCEPT	
defined .....	68, 84, 91
spaces .....	73
concept definition	
precision .....	137
recall .....	137
concept matching	
preference .....	67
Concept menu	
Add Concept .....	10
Create Directory Tree .....	11
Delete All Selected Concepts .....	10
Delete Concept .....	10
Priorities .....	11
Rename Concept .....	10
concept node	
drop-down selections .....	48
Concept Syntax Check window	
usage .....	41
CONCEPT_RULE	
defined .....	68, 98, 100, 102, 104, 111
spaces .....	73
Concepts node	
drop-down menu .....	47
concordance view	
TEST button .....	23
Concordance window	
Insert text markers .....	30
Project Settings .....	26, 29, 61
Show Filename .....	30
Show Full Path .....	30
Test multiple files .....	30, 163
usage .....	22
concordance window	30
access .....	18, 22
For each match show .....	29
Hide Filenames .....	30
Insert text markers .....	163
match displays .....	29
Sort by .....	30

---

Copy All Selections	
Edit menu .....	8
coref	
CLASSIFIER .....	118
coreference	
canonical form .....	114
operators .....	114
Create Directory Tree	
Concept menu .....	11
Create Folders option	
defined .....	20
curly braces	
usage .....	72
Cut All Selections	
Edit menu .....	8

## D

Data tab	
concepts .....	19
defined .....	6, 16
Data window	
Priority field .....	75, 76, 105
Decrease Font Size	
Testing menu .....	12
definition	
analyze .....	158
Definition tab	
defined .....	6, 16
Delete All Selected Concepts	
Concept menu .....	10
Delete Concept	
Concept menu .....	10
defined .....	48
usage .....	48
Delete Language	
Project menu .....	10
usage .....	46
Delete Selected Test File	
Testing menu .....	11

---

dictionary entries	
part-of-speech tags .....	96
disambiguation	
defined .....	89
DIST	
defined .....	79
usage .....	80, 102, 104
document	
PARA .....	67
SENT .....	67
Document tab	
defined .....	6, 16
Document window	
Fact .....	110
usage .....	21, 24, 144
Web browser .....	23
duplicate instances	
return .....	86

## E

Edit menu	
Copy All Selections .....	8
Cut All Selections .....	8
Find in All Rules .....	9
Paste .....	8
Paste Single Node .....	8
Paste Symbolic Link .....	8
Text Find .....	8
Text Replace .....	8
Tree Find .....	8
Tree Replace .....	8
Enable Concepts	
Project menu .....	10
usage .....	46
Enter Display Name	
Enter Names window .....	34
Enter name for internal data files	
Enter Names window .....	34
Enter Names window	
access .....	33

---

Enter Display Name .....	34
Enter name for internal data files .....	34
usage .....	33
Use same name for both fields .....	33
existing project	
access .....	50
Exit option	
File menu .....	7
Expand Full	
usage .....	46
Expand Fully	
usage .....	46, 47, 48
Expand Fully option	
usage .....	48
export feature	
usage .....	91, 92
export symbol	
_ref .....	119
exported terms	
not in rule .....	92

## F

Fact	
defined .....	107
Document window .....	110
multiple .....	109
fact .....	109
argument .....	113
defined .....	107
PREDICATE_RULE .....	109
view matches .....	108
failing documents	
defined .....	147
File menu	
New Project .....	7
Open Project .....	7
Save Project .....	7
Save Project As .....	7
Find field	
Tree Find window .....	37

---

Find in All Rules	
Edit menu .....	9
Find Next button	
Tree Find window .....	37
For each match show	
concordance window .....	29
Forward button	
Browser .....	12
defined .....	25

## G

Go button	
defined .....	24

## H

Hide Filenames	
concordance window .....	30
Home button	
Browser .....	12
defined .....	25

## I

icons	
Standard toolbar .....	13
Identical Path Name option	
defined .....	20
Import Failing Test Files	
Testing menu .....	11
Import Test Files	
Testing menu .....	11
Import Test Files option	
usage .....	139
In-concept files	
usage .....	149
Increase Font Size	
Testing menu .....	12

---

Insert text markers	
Concordance window .....	30
concordance window .....	163
interface	
view .....	5

## L

language fonts	
UTF-8 encoding .....	53
language node	
drop-down menu .....	46
LITI window	
Project Settings .....	26
Ln	
defined .....	17
Load Text	
defined .....	17
location	
matches occur .....	121
logical operators	
table .....	79
Longest	
Data window .....	106
usage .....	26, 60, 110

## M

main window	
components .....	6
Match case option	
Tree Find window .....	37
Menu bar	
defined .....	6
Misc tab	
Project Settings .....	121
XML Tags to Ignore .....	28
Misc window	
Compatibility Date .....	28
Paragraph Separator .....	28

---

Project Settings .....	26
usage .....	27
multiple rules	
add .....	123

## N

navigating concepts .....	63
New Project	
File menu .....	7
New Project window	
access .....	7
NO_BREAK	
defined .....	68, 88
usage .....	88
Number of Taxonomy Nodes	
View menu .....	9
Number of Taxonomy Nodes option	
usage .....	34
Number of Taxonomy Nodes window	
usage .....	34

## O

Open Project	
File menu .....	7
Open Test Document	
Testing menu .....	11
Open window	
access .....	7
OR	
defined .....	79
ORDDIST	
defined .....	79
usage .....	81, 104
Out-of-concept files	
usage .....	149
Overlapping Concept Matches	
usage .....	26, 60

---

## P

PARA	
document	67
paragraph field	
match	122
Paragraph Separator	
Misc window	28
parentheses	
usage	72
partial match	95
part-of-speech tags	
codes	174
requirements	96
Paste	
Edit menu	8
Paste Single Node	
Edit menu	8
usage	48
Paste Symbolic Link	
Edit menu	8
percent	
REGEX	97
usage	75
precision	
concept definition	137
PREDICATE_RULE	
defined	69, 109
fact	109
Prep	
defined	96
Priorities	
Concept menu	11
priority	
overlapping matches	94
rank	120
usage	75
Priority field	
Data window	75, 76, 105
defined	20
definition	94

---

PRIORITY specification	
usage .....	119
Program and Project title bar	
defined .....	6
Project menu	
Add Language .....	9
Delete Language .....	10
Enable Concepts .....	10
Remove Concepts .....	10
Settings .....	10
project name node	
drop-down menu .....	45
Project Settings	
Concordance window .....	26, 29, 61
LITI window .....	26
Misc tab .....	121
Misc window .....	26
Priority field .....	94
REMOVE_ITEM .....	27
Project Settings - LITI tab	
rule matches .....	67
Project Settings windows	
usage .....	25
Propagate Button	
usage .....	20
Propagate Options	
defined .....	20

## Q

quotation marks	
usage .....	72

## R

rank	
priority .....	120
recall	
concept definition .....	137

---

Refresh button	
Browser .....	12
defined .....	25
Refresh Tree	
View menu .....	9
Refresh Tree button	
usage .....	156
REGEX	
defined .....	68, 97
percent .....	97
usage .....	104
regular expressions	
usage .....	75
relative rankings	
increase .....	94
Remove Concepts	
Project menu .....	10
usage .....	47
Remove duplicate facts	
usage .....	106, 113
Remove Tags	
Testing menu .....	12
Remove Tags option	
defined .....	25
REMOVE_ITEM	
defined .....	68, 89
Project Settings .....	27
Rename Concept	
Concept menu .....	10
usage .....	48
Replace button	
Tree Find window .....	38
Return all identical matches	
Data window .....	106
usage .....	27, 60, 111
rule matches	
Priority Settings .....	67
rules	
defined .....	67

---

## S

Save Project	
File menu .....	7
Save Project As	
File menu .....	7
Save Test Document	
Testing menu .....	11
Select a Language window	
usage .....	31, 45
Selected concept	
define .....	21
defined .....	25
test against .....	25
SENT	
defined .....	79
document .....	67
usage .....	81
SENT_n	
defined .....	79
usage .....	81
SENTEND_n	
defined .....	79
usage .....	82
SENTSTART_n	
usage .....	81
sep	
defined .....	96
usage .....	123
SEQUENCE	
defined .....	69, 107
usage .....	123
Settings	
Project menu .....	10
Show Filename	
Concordance window .....	30
Show Full Path	
Concordance window .....	30
Sort by .....	30
concordance window .....	30
square braces	
usage .....	72

---

Standard toolbar	
icons .....	13
standard toolbar	
defined .....	6
Start	
menu .....	5
Status Bar	
usage .....	13
status window	
defined .....	24
Stop button	
Browser .....	12
defined .....	24, 25
subnodes	
count .....	35
Syntax Check button	
defined .....	17

## T

Taxonomy as Text	
View menu .....	9
Taxonomy tab	
defined .....	6
taxonomy window	
navigation within .....	63
TEST button	
defined .....	24
Test Disabled field	
defined .....	20
Test File column	
heading .....	149
Test File field	
defined .....	24
Test File window	
usage .....	139
Test files for this concept	
select .....	148
Test multiple files	
Concordance window .....	30, 163

---

Testing menu	
Browser .....	12
Clear Test Document .....	11
Decrease Font Size .....	12
Delete Selected Test File .....	11
Import Failing Test Files .....	11
Import Test Files .....	11
Increase Font Size .....	12
Open Test Document .....	11
Remove Tags .....	12
Save Test Document .....	11
Testing Path field	
defined .....	20
testing process	
customize .....	146
testing set	
UTF-8 encoding .....	53
Testing tab	
defined .....	6, 16
Testing window	
testing results .....	158
usage .....	144
Text Find	
Edit menu .....	8
Text Replace	
Edit menu .....	8
token	
defined .....	70
usage .....	98, 104
Tree Find	
Edit menu .....	8
Tree Find icon	
usage .....	37
Tree Find window	
access .....	37
Find field .....	37
Find Next button .....	37
Match case option .....	37
Replace button .....	38
usage .....	36
Tree Replace	
Edit menu .....	8

---

Tree Replace window	
usage .....	38

## U

Use same name for both fields	
Enter Names window .....	33
user interface	
display .....	5
UTF-8 encoding	
language fonts .....	53
testing set .....	53

## V

View menu	
Number of Taxonomy Nodes .....	9
Refresh Tree .....	9
Taxonomy as Text .....	9

## W

Web browser	
Document window .....	23

## X

XML Tags to Ignore	
Misc tab .....	28

