

IBM SPSS Text Analytics for Surveys 4.0.1 User's Guide



Note: Before using this information and the product it supports, read the general information under “Notices” on p. 255.

This edition applies to IBM® SPSS® Text Analytics for Surveys 4 and to all subsequent releases and modifications until otherwise indicated in new editions.

Adobe product screenshot(s) reprinted with permission from Adobe Systems Incorporated.

Microsoft product screenshot(s) reprinted with permission from Microsoft Corporation.

Licensed Materials - Property of IBM

© Copyright IBM Corporation 2004, 2011.

U.S. Government Users Restricted Rights - Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

Preface

Welcome to IBM® SPSS® Text Analytics for Surveys version 4.0.1, a survey text coding application that provides for meaningful analysis of responses to open-ended questions. With this product, anyone performing survey research can quickly transform unstructured survey responses into quantitative data. “Unlocking” this open-ended text data can significantly improve analysis quality and decision-making ability. This application allows you to import survey data, extract key concepts, refine the results, and categorize responses. Once you have categorized your data, you can export your categories for import into quantitative analytic tools, such as the IBM® SPSS® Statistics system, for further analysis and graphing.

SPSS Text Analytics for Surveys combines advanced linguistic technologies designed to reliably extract and classify key concepts within open-ended survey responses with manual techniques. Using robust category building algorithms and simple drag-and-drop functionality, you can create categories, or “codes,” into which your survey responses will be categorized. The categories produced can also be reused to provide consistent results across the same or similar studies. Since open-ended response data can vary immensely from one survey to another, no two projects will be exactly the same; however, you can expect to follow the same basic process to accomplish your analysis. For more information, see the topic “The Typical Process” in Chapter 2 on p. 9.

About IBM Business Analytics

IBM Business Analytics software delivers complete, consistent and accurate information that decision-makers trust to improve business performance. A comprehensive portfolio of business intelligence, predictive analytics, financial performance and strategy management, and analytic applications provides clear, immediate and actionable insights into current performance and the ability to predict future outcomes. Combined with rich industry solutions, proven practices and professional services, organizations of every size can drive the highest productivity, confidently automate decisions and deliver better results.

As part of this portfolio, IBM SPSS Predictive Analytics software helps organizations predict future events and proactively act upon that insight to drive better business outcomes. Commercial, government and academic customers worldwide rely on IBM SPSS technology as a competitive advantage in attracting, retaining and growing customers, while reducing fraud and mitigating risk. By incorporating IBM SPSS software into their daily operations, organizations become predictive enterprises – able to direct and automate decisions to meet business goals and achieve measurable competitive advantage. For further information or to reach a representative visit <http://www.ibm.com/spss>.

Technical support

Technical support is available to maintenance customers. Customers may contact Technical Support for assistance in using IBM Corp. products or for installation help for one of the supported hardware environments. To reach Technical Support, see the IBM Corp. web site at <http://www.ibm.com/support>. Be prepared to identify yourself, your organization, and your support agreement when requesting assistance.

Contents

Part I: Getting Started

1 About Text Analysis 1

What's New	1
Open-Ended Survey Data	2
About Text Mining.	3
How Extraction Works.	3
How Categorization Works.	6
Preparing for Text Analysis.	7
Reliability and Fine-Tuning	8
Refining Linguistic Resources	8
Refining Category Definitions.	8

2 Getting Started 9

The Typical Process	9
The Text Analysis Window	10
The Question View.	11
The Entire Project View	13
The Resource Editor Window	14
Setting Options.	16
Options: System Tab	17
Options: Display Tab	18
Options: Sounds Tab	20
Options: Translation Tab	21
Microsoft Internet Explorer Settings for Help	22

Part II: Text Analysis

3 Creating Projects and Packages 25

Creating Projects	25
Preparing Your Data	26

Starting New Projects	27
Selecting Data Sources	28
Selecting Variables	32
Translating into English	34
Selecting Categories and Resources	36
Using Text Analysis Packages.	40
Making Text Analysis Packages.	40
Updating Text Analysis Packages	42

4 Working with Projects 47

Opening Projects	47
Editing Project Properties	48
Viewing Project Data	49
Sorting Variables	50
Editing Variable Properties.	50
Saving Projects	51
Exporting Categorization Results.	52
Exporting to IBM SPSS Statistics or IBM SPSS Data Collection	54
Exporting to Microsoft Excel	56
Exporting Summary Graphs.	58
Changing Data Sources	61
Selecting Data Sources	62
Selecting Variables	66
Matching Variables	68
Translating into English	69
Updating Data.	71
Translating into English.	71
Sharing Projects.	73
Flagging Responses	73
Project Status Bar	74

5 Extracting Data 77

Extracted Results: Concepts, Types, and Patterns.	77
Extracting Data.	81
Saving Extraction Results	84

Refining Extraction Results	84
Adding Synonyms	85
Adding Concepts to Types	87
Excluding Concepts from Extraction	89
Forcing Words into Extraction	90

6 *Categorizing Text Data* **91**

The Categories Pane	92
The Data Pane	95
Category Relevance	97
Methods and Strategies for Creating Categories	98
Methods for Creating Categories	98
Strategies for Creating Categories	99
Tips for Creating Categories	100
Choosing the Best Descriptors	101
About Categories	103
Category Properties	104
Building Categories	105
Advanced Linguistic Settings	109
About Linguistic Techniques	113
Advanced Frequency Settings	118
Extending Categories	120
Creating Categories Manually	124
Creating New or Renaming Categories	124
Creating Categories by Drag-and-Drop	125
Importing and Exporting Predefined Categories	126
Importing Predefined Categories	127
Exporting Categories	135
Using Category Rules	138
Category Rule Syntax	139
Using TLA Patterns in Category Rules	140
Using Wildcards in Category Rules	142
Category Rule Examples	144
Creating Category Rules	146
Editing and Deleting Rules	147
Editing and Refining Categories	148
Editing Category Properties	149
Adding Descriptors to Categories	150
Editing Category Descriptors	150
Moving Categories	151

Merging or Combining Categories	152
Forcing Responses into Categories	153
Text Matching in Categories	154
Copying Categories	156
Printing Categories	156
Deleting Categories	157

7 Visualizing Graphs 159

Category Bar Chart	160
Category Web Graph	161
Category Web Table	162
Using Graph Toolbars and Palettes	162
Editing Visualizations	163
General Rules for Editing Visualizations	164
Editing and Formatting Text	165
Changing Colors, Patterns, Dashings, and Transparency	166
Rotating and Changing the Shape and Aspect Ratio of Point Elements	167
Changing the Size of Graphic Elements	167
Specifying Margins and Padding	168
Formatting Numbers	169
Changing the Axis and Scale Settings	170
Editing Categories	171
Changing the Orientation Panels	173
Transforming the Coordinate System	173
Changing Statistics and Graphic Elements	175
Changing the Position of the Legend	178
Copying a Visualization and Visualization Data	178
Keyboard Shortcuts	179

Part III: Resource Editor

8 Templates and Resources 183

The Editor Interface	184
Making and Updating Templates	186
Switching Resource Templates	187
Managing Templates	188

Importing and Exporting Templates	189
Backing Up Resources	190
Importing Resource Files	192

9 Working with Libraries 195

Shipped Libraries	195
Creating Libraries	196
Adding Public Libraries	197
Finding Terms and Types	198
Viewing Libraries	199
Managing Local Libraries	199
Renaming Local Libraries	199
Disabling Local Libraries	200
Deleting Local Libraries	200
Managing Public Libraries	201
Sharing Libraries	202
Publishing Libraries	204
Updating Libraries	204
Resolving Conflicts	205

10 About Library Dictionaries 207

Type Dictionaries	207
Built-in Types	208
Creating Types	209
Adding Terms	210
Forcing Terms	214
Renaming Types	215
Moving Types	216
Disabling and Deleting Types	216
Substitution/Synonym Dictionaries	217
Defining Synonyms	218
Defining Optional Elements	220
Disabling and Deleting Substitutions	221
Exclude Dictionaries	222

11 About Advanced Resources 225

Finding	226
Replacing	226
Fuzzy Grouping	227
Nonlinguistic Entities	228
Regular Expression Definitions	229
Normalization	231
Configuration	232
Language Handling	233
Extraction Patterns	233
Forced Definitions	234
Abbreviations	234

Appendices

A Japanese Text Exceptions 237

Extracting and Categorizing Japanese Text	237
How Extraction Works	237
How Secondary Extraction Works	239
How Categorization Works	242
Editing Resources for Japanese Text	242
Japanese Library Tree, Types, and Term Pane	244
Available Types for Japanese Text	246
Editing Japanese Type Properties	250
Using the Synonym Dictionary for Japanese Text	250
Validating and Compiling Japanese Resources	252
Other Exceptions for Japanese	252

B Notices 255

Index 257

Part I:
Getting Started

About Text Analysis

Welcome to IBM® SPSS® Text Analytics for Surveys version 4.0.1, a survey text coding application that provides for meaningful analysis of responses to open-ended questions. With this product, anyone performing survey research can quickly transform unstructured survey responses into quantitative data. “Unlocking” this open-ended text data can significantly improve analysis quality and decision-making ability. This application allows you to import survey data, extract key concepts, refine the results, and categorize responses. Once you have categorized your data, you can export your categories for import into quantitative analytic tools, such as the IBM® SPSS® Statistics system, for further analysis and graphing.

SPSS Text Analytics for Surveys combines advanced linguistic technologies designed to reliably extract and classify key concepts within open-ended survey responses with manual techniques. Using robust category building algorithms and simple drag-and-drop functionality, you can create categories, or “codes,” into which your survey responses will be categorized. The categories produced can also be reused to provide consistent results across the same or similar studies. Since open-ended response data can vary immensely from one survey to another, no two projects will be exactly the same; however, you can expect to follow the same basic process to accomplish your analysis. For more information, see the topic “The Typical Process” in Chapter 2 on p. 9.

What’s New

The following new features can be found in IBM® SPSS® Text Analytics for Surveys 4.0.1:

Hierarchical categories. Categories can now have a hierarchical structure, meaning they can contain subcategories and those subcategories can also have subcategories of their own and so on. You can import predefined category structures, formerly called code frames, with hierarchical categories as well as build these hierarchical categories inside the product.

In effect, hierarchical categories enable you to build a tree structure with one or more subcategories to group items such as different concept or topic areas more accurately. A simple example can be related to leisure activities; answering a question such as *What activity would you like to do if you had more time?* you may have top categories such as *sports, art and craft, fishing*, and so on; down a level, below *sports*, you may have subcategories to see if this is *ball games, water-related*, and so on.

Language Weaver access. The way you access the Language Weaver translation interface has been simplified to use a single URL and associated security details.

Open-Ended Survey Data

Survey questionnaires commonly contain different kinds of questions, including open- and closed-ended questions. The first, a **closed-ended question**, presents a limited set of responses that allows for various types of quantitative analyses. The second, an **open-ended question**, permits a respondent to provide an unstructured response of varying length and detail.

The words people use to answer a question tell you a lot about what they think and feel. That is why open-ended questions are often included in surveys: they provide more varied and textured information than closed-ended questions and often can provide insight that was not anticipated by the survey designer.

However, the use of a larger number of open-ended questions has traditionally been viewed as cost-prohibitive because of the analytical overhead incurred with the interpretation of responses. Furthermore, these long responses must be coded in a standardized manner, using a detailed set of coding instructions. This task can be difficult because coders often disagree on how to categorize specific responses. When coders disagree, the reliability of the results is reduced. For all these reasons, the coding of open-ended responses has long been viewed as time-consuming and expensive, often outweighing the benefits derived from the data collected.

IBM® SPSS® Text Analytics for Surveys offers an alternative to this costly procedure, since it can accomplish the coding of open-ended responses in a fraction of the time required to do the job manually. Through the use of advanced linguistic theory and technologies, SPSS Text Analytics for Surveys analyzes open-ended response text as a set of phrases and sentences whose grammatical structure provides context for the meaning of a response. After analyzing this text, the key concepts and word patterns are extracted and classified into categories.

You can use built-in category building techniques to automatically create categories and manual techniques to fine-tune the results. The reliability of results increases dramatically, since extraction and categorization are always performed in a consistent and repeatable manner—the same response is categorized in the same categories every time unless you choose to fine-tune your category definitions or linguistic libraries.

Successful survey analysis does not depend on one approach alone. The subjective nature of open-ended response interpretation calls for the use of multiple techniques. In addition to its built-in extraction and category building techniques, SPSS Text Analytics for Surveys also relies on the user's grasp of the specific text analysis goals for each survey. Text analysis is most powerful when performed in an iterative manner (extract, review, refine, reextract), and its usefulness will often depend on the amount of time and effort spent manually reviewing and refining extraction results and category definitions. For more information, see the topic "Reliability and Fine-Tuning" on p. 8.

If you work with identical or similar questions in reoccurring surveys, you can reuse categories in other questions or projects. Reusing categories allows for much greater consistency in coding, as well as offering a huge savings in time and effort.

Additionally, you may want to perform other analyses. The categories you produce can be used in various types of statistical analyses with the other questions in the questionnaire or other demographic data to gain further insight into the respondents and their opinions and behaviors. After using SPSS Text Analytics for Surveys to discover the categories that underlie a set of responses, you can also export the categories for further quantitative analysis in another program such as IBM® SPSS® Statistics Base.

About Text Mining

Text analysis, a form of qualitative analysis, is the extraction of useful information from text (such as open-ended responses) so that the key ideas or concepts contained within this text can be grouped into an appropriate number of categories. Text analysis can be performed on all types and lengths of text, although the approach to the analysis will vary somewhat.

Shorter records are most easily categorized, since they are not as complex and usually contain fewer ambiguous words and responses. For example, with short, open-ended survey questions, if we ask people to name their three favorite vacation activities, we might expect to see many short answers, such as *going to the beach*, *visiting national parks*, or *doing nothing*. Longer, open-ended responses, on the other hand, can be quite complex and very lengthy, especially if respondents are educated, motivated, and have enough time to complete a questionnaire. If we ask people to tell us about their political beliefs in a survey or have a blog feed about politics, we might expect some lengthy comments about all sorts of issues and positions.

Survey researchers do not normally analyze very long responses. Responses on most surveys tend to be short to medium in length (a sentence to a short paragraph). IBM® SPSS® Text Analytics for Surveys was designed to handle this length of text but can analyze responses that are much longer.

There are several different methods of text analysis. First, there is the manual approach: having people read the survey responses, note their contents, determine the key concepts they contain, and assign codes to them. Because people are good at understanding text, this approach is quite accurate. But it is time-consuming, labor-intensive and, with the immense volume of text now available, increasingly impractical. This approach also relies heavily on the interpretation of each coder.

A different approach is to employ automated solutions. There are many different automated solutions to choose from, including statistical and linguistic solutions. SPSS Text Analytics for Surveys offers a combination of automated linguistic and statistical techniques to yield the most reliable results for each stage of the process. In this product, linguistic-based techniques are used to extract the key concepts from the responses automatically, and both linguistic and statistical techniques can be used to create the category definitions (codes) that are assigned to responses.

How Extraction Works

During the extraction of key concepts and ideas from your responses, IBM® SPSS® Text Analytics for Surveys relies on linguistics-based text analysis. This approach offers the speed and cost effectiveness of statistics-based systems. But it offers a far higher degree of accuracy, while requiring far less human intervention. Linguistics-based text analysis is based on the field of study known as natural language processing, also known as computational linguistics.

To illustrate the difference between statistics-based and linguistics-based approaches during the extraction process, consider how each would respond to a query about reproduction of documents. Both statistics-based and linguistics-based solutions would have to expand the word reproduction to include synonyms, such as copy and duplication. Otherwise, relevant information will be overlooked. But if a statistics-based solution attempts to do this type of synonymy—searching for other terms with the same meaning—it is likely to include the term birth as well, generating a number of irrelevant results. The understanding of language cuts

through the ambiguity of text, making linguistics-based text mining, by definition, the more reliable approach.

Understanding how the extraction process works can help you make key decisions when fine-tuning your linguistic resources (libraries, types, synonyms, and more). Steps in the extraction process include:

- Converting source data to a standard format
- Identifying candidate terms
- Identifying equivalence classes and integration of synonyms
- Assigning a type
- Indexing
- Matching patterns and events extraction

Step 1. Converting source data to a standard format

In this first step, the data you import is converted to a uniform format that can be used for further analysis. This conversion is performed internally and does not change your original data.

Step 2. Identifying candidate terms

It is important to understand the role of linguistic resources in the identification of candidate terms during linguistic extraction. Linguistic resources are used every time an extraction is run. They exist in the form of templates, libraries, and compiled resources. Libraries include lists of words, relationships, and other information used to specify or tune the extraction. The compiled resources cannot be viewed or edited. However, the remaining resources (templates) can be edited in the Resource Editor.

Compiled resources are core, internal components of the extraction engine within IBM® SPSS® Text Analytics for Surveys. These resources include a general dictionary containing a list of base forms with a part-of-speech code (noun, verb, adjective, adverb, participle, coordinator, determiner, or preposition). The resources also include reserved, built-in types used to assign many extracted terms to the following types, <Location>, <Organization>, or <Person>. For more information, see the topic “Built-in Types” in Chapter 10 on p. 208.

In addition to those compiled resources, several libraries are delivered with the product and can be used to complement the types and concept definitions in the compiled resources, as well as to offer other types and synonyms. These libraries—and any custom ones you create—are made up of several dictionaries. These include type dictionaries, substitution dictionaries (synonyms and optional elements), and exclude dictionaries. For more information, see the topic “Working with Libraries” in Chapter 9 on p. 195.

Once the data have been imported and converted, the extraction engine will begin identifying candidate terms for extraction. Candidate terms are words or groups of words that are used to identify concepts in the text. During the processing of the text, single words (**uniterms**) that are not in the compiled resources are considered as candidate term extractions. Candidate compound words (**multiterms**) are identified using part-of-speech pattern extractors. For example, the multiterm `sports car`, which follows the “adjective noun” part-of-speech pattern, has two components. The multiterm `fast sports car`, which follows the “adjective adjective noun” part-of-speech pattern, has three components.

Note: The terms in the aforementioned compiled general dictionary represent a list of all of the words that are likely to be uninteresting or linguistically ambiguous as uniterms. These words are excluded from extraction when you are identifying the uniterms. However, they are reevaluated when you are determining parts of speech or looking at longer candidate compound words (multiterms).

Finally, a special algorithm is used to handle uppercase letter strings, such as job titles, so that these special patterns can be extracted.

Step 3. Identifying equivalence classes and integration of synonyms

After candidate uniterms and multiterms are identified, the software uses a set of algorithms to compare them and identify equivalence classes. An equivalence class is a base form of a phrase or a single form of two variants of the same phrase. The purpose of assigning phrases to equivalence classes is to ensure that, for example, `president of the company` and `company president` are not treated as separate concepts. To determine which concept to use for the equivalence class—that is, whether `president of the company` or `company president` is used as the lead term, the extraction engine applies the following rules in the order listed:

- The user-specified form in a library.
- The most frequent form in the full body of text.
- The shortest form in the full body of text (which usually corresponds to the base form).

Step 4. Assigning type

Next, types are assigned to extracted concepts. A type is a semantic grouping of concepts. Both compiled resources and the libraries are used in this step. Types include such things as higher-level concepts, positive and negative words, first names, places, organizations, and more. Additional types can be defined by the user. For more information, see the topic “Type Dictionaries” in Chapter 10 on p. 207.

Step 5. Indexing

The entire set of records is indexed by establishing a pointer between a text position and the representative term for each equivalence class. This assumes that all of the inflected form instances of a candidate concept are indexed as a candidate base form. The global frequency is calculated for each base form.

Step 6. Matching patterns and events extraction

IBM SPSS Text Analytics for Surveys can discover not only types and concepts but also relationships among them. Several algorithms and libraries are available with this product and provide the ability to extract relationship patterns between types and concepts. They are particularly useful when attempting to discover specific opinions (for example, product reactions) or the relational links between people or objects (for example, links between political groups or genomes).

How Categorization Works

There are several different techniques you can choose to create categories. Because every dataset is unique, the number of techniques and the order in which you apply them may change. Since your interpretation of the results may be different from someone else's, you may need to experiment with the different techniques to see which one produces the best results for your text data.

In this guide, **category building** refers to the generation of category definitions and classification through the use of one or more built-in techniques, and **categorization** refers to the scoring, or labeling, process whereby unique identifiers (name/ID/value) are assigned to the category definitions for each record.

During category building, the concepts and types that were extracted are used as the building blocks for your categories. When you build categories, the records are automatically assigned to categories if they contain text that matches an element of a category's definition.

IBM® SPSS® Text Analytics for Surveys offers you several automated category building techniques to help you categorize your records quickly.

Grouping Techniques

Each of the techniques available is well suited to certain types of data and situations, but often it is helpful to combine techniques in the same analysis to capture the full range of records. You may see a concept in multiple categories or find redundant categories.

Concept Root Derivation. This technique creates categories by taking a concept and finding other concepts that are related to it by analyzing whether any of the concept components are morphologically related, or share roots. This technique is very useful for identifying synonymous compound word concepts, since the concepts in each category generated are synonyms or closely related in meaning. It works with data of varying lengths and generates a smaller number of compact categories. For example, the concept `opportunities to advance` would be grouped with the concepts `opportunity for advancement` and `advancement opportunity`. For more information, see the topic "Concept Root Derivation" in Chapter 6 on p. 114.

Semantic Network. This technique begins by identifying the possible senses of each concept from its extensive index of word relationships and then creates categories by grouping related concepts. This technique is best when the concepts are known to the semantic network and are not too ambiguous. It is less helpful when text contains specialized terminology or jargon unknown to the network. In one example, the concept `granny smith apple` could be grouped with `gala apple` and `winesap apple` since they are siblings of the `granny smith`. In another example, the concept `animal` might be grouped with `cat` and `kangaroo` since they are hyponyms of `animal`. This technique is available for English text only in this release. For more information, see the topic "Semantic Networks" in Chapter 6 on p. 116.

Concept Inclusion. This technique builds categories by grouping multiterm concepts (compound words) based on whether they contain words that are subsets or supersets of a word in the other. For example, the concept `seat` would be grouped with `safety seat`, `seat belt`, and `seat belt buckle`. For more information, see the topic "Concept Inclusion" in Chapter 6 on p. 115.

Co-occurrence. This technique creates categories from co-occurrences found in the text. The idea is that when concepts or concept patterns are often found together in documents and records, that co-occurrence reflects an underlying relationship that is probably of value in your

category definitions. When words co-occur significantly, a co-occurrence rule is created and can be used as a category descriptor for a new subcategory. For example, if many records contain the words `price` and `availability` (but few records contain one without the other), then these concepts could be grouped into a co-occurrence rule, (`price & available`) and assigned to a subcategory of the category `price` for instance. For more information, see the topic “Co-occurrence Rules” in Chapter 6 on p. 117.

- **Minimum number of records.** To help determine how interesting co-occurrences are, define the minimum number of records that must contain a given co-occurrence for it to be used as a descriptor in a category.

Preparing for Text Analysis

Text analysis involves more than extraction and categorization. To successfully analyze text, consider the following points:

- As in survey design, the quality of the responses you import into IBM® SPSS® Text Analytics for Surveys directly affects the quality of the resulting categorizations. In general, vague or unclear questions result in responses that can drift and wander and be quite difficult to analyze.
- Like statistical analysis, text analysis should be performed with clear objectives in mind. Before you begin any analysis, you should reflect on your study and determine what it is that you are trying to learn.

For example, let’s assume that a survey was conducted in a local school district to measure parents’ attitudes regarding the quality of education their children have received. During the analysis, we could focus on topics such as teacher names, school programs, and so on, or we could focus on identifying and grouping positive feedback and negative feedback. Likewise, the level of granularity required for the analysis must be defined, such as grouping all remarks about funding together or breaking this category down further into funding per program. The codes, or categories, we create should reflect the focus and objectives of our analysis.

- Far more than statistical analysis, text analysis is not an exact science, since there is no one “correct” outcome. Text analysis is performed with objectives in mind, but it is also subjective in that it is influenced by the analyst’s interpretation of the message conveyed by the respondent—for example, how to identify and classify attitudes filled with sarcasm. Depending on their objectives and focus, two competent people can analyze the same data and reach different conclusions.
- Text analysis is very much an *iterative process*. As you work with your survey responses, you will surely reextract and recategorize your responses using different category definitions (that is, coding schemes), different concept or synonym definitions, and different groupings of responses.

After you have extracted concepts from your text and created your categories, you should review your results carefully. If you find any elements you want to tweak, simply adjust your analysis by fine-tuning your category definitions and linguistic library definitions. Then the responses will automatically be recategorized when you reextract. You may go through this process one or many times before you are satisfied with the results of your analysis.

Note: For more information on considerations before importing data, see “Preparing Your Data” on p. 26

Reliability and Fine-Tuning

Whenever you code data, you want the resulting categories to be reliable. In the context of coding open-ended responses, this means that two independent coders, using the same rules (coding frame), will code the same response identically. When text analysis is done manually, this is a critical issue. A valuable set of categories can be created, but if they cannot be reliably applied to the responses, their value decreases substantially. When IBM® SPSS® Text Analytics for Surveys is applied to the same data, with the same linguistic resources, it will always reproduce a prior analysis perfectly. It is 100% reliable.

However, this does not mean that there will be no errors in the analysis, but the focus on coding can now shift to something else—*fine-tuning*. In human coding, the coders read the response and can capture all of the nuances of a statement (even if they have trouble applying the coding categories). SPSS Text Analytics for Surveys can apply the coding categories, but the categories have to be defined so that nuances and distinctions are captured. There are two ways that fine-tuning can be performed:

- Refining the linguistic resources
- Refining the category definitions

Refining Linguistic Resources

IBM® SPSS® Text Analytics for Surveys will easily create categories with no intervention on your part, but they will invariably not capture all of the information in the responses. You need to work to improve the linguistic base that the program uses so that its category creation becomes more and more tuned to the idiosyncrasies of the text. To improve this base, you can customize and fine-tune the linguistic resources used in extracting from the text.

Fine-tuning, in this case, involves adding words and phrases to various linguistic libraries and dictionaries, specifying words to be excluded from the analysis, defining synonyms, or creating custom libraries with a specific goal in mind. This goal is to accurately capture the ideas of the respondents in the text and remove ambiguity in the results.

Refining Category Definitions

In addition to refining the linguistic resources, you should review your categories by looking for ways to combine or clean up their definitions as well as checking some of the categorized responses. You can use the automated category building techniques to create your categories; however, you will surely want to perform a few tweaks to these definitions. After using a technique, a number of new categories appear in the window. You can expand the categories to see the concepts that define each category. You can then review the responses in a category and make adjustments until you are comfortable with your category definitions.

None of the automatic techniques will perfectly categorize your data; therefore we recommend finding and applying one or more automatic techniques that work well with your data. You can then use manual techniques to make minor adjustments, remove any misclassifications, or add records or concepts that may have been missed.

Getting Started

This documentation presents the tasks that you can perform with IBM® SPSS® Text Analytics for Surveys and the techniques that you can use to categorize your responses. The information presented here guides you through your initial analysis. It discusses all of the processes to fully analyze your data, but because every data set is different, you will ultimately need to decide when your analysis is complete.

In this chapter, we discuss the typical process users go through when performing text analysis. The interface is also explained from a high-level perspective along with the major tasks and elements you will work with.

The Typical Process

The following is a summary of the typical work flow process that you will follow while using IBM® SPSS® Text Analytics for Surveys.

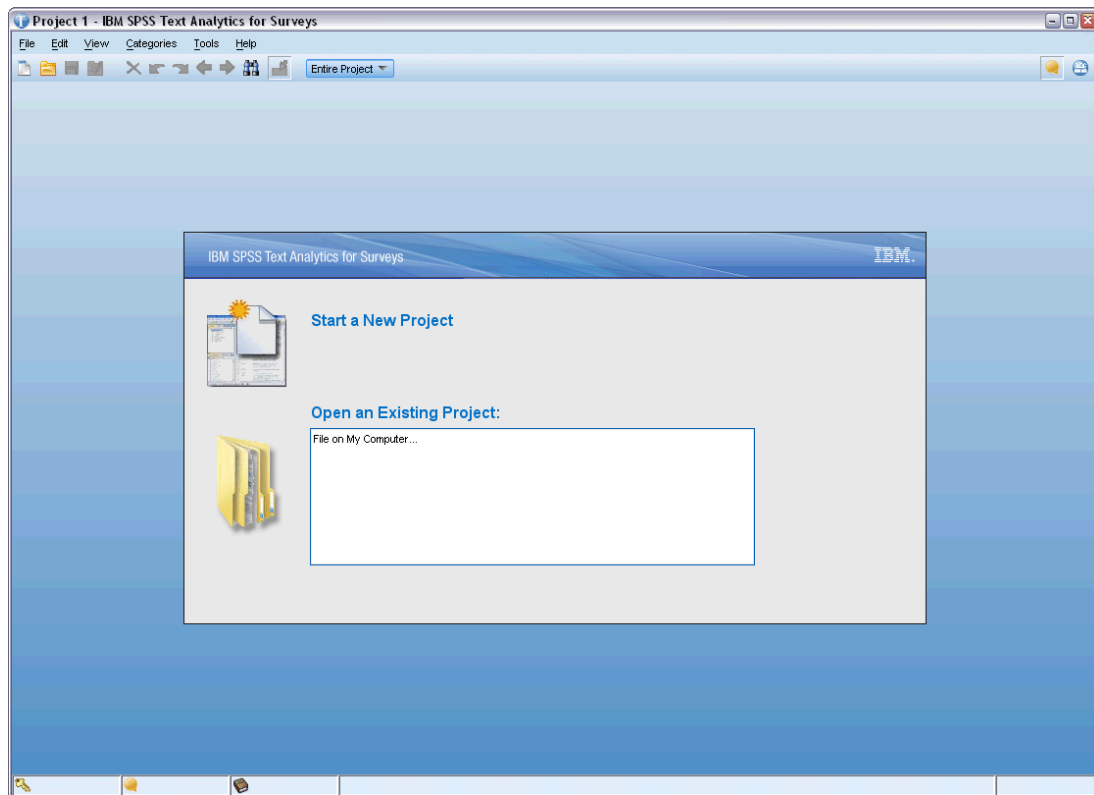
- **Create a project** by importing your survey data, including open-ended response(s), an ID variable, and other reference variables into SPSS Text Analytics for Surveys. Data can be read from IBM® SPSS® Statistics data files, Microsoft Excel, any ODBC-compliant database program, or an IBM® SPSS® Data Collection data source. You can choose a text analysis package to benefit from some predefined categories and specialized resources to get up and running quickly.
- **Extract** concepts and patterns for each open-ended question you imported. The internal extraction engine automatically identifies and collects the key terms expressed in the text. These terms are grouped under a main concept. Concepts are then grouped into types, which are collections of similar terms, such as organizations, products, or positive opinions. Patterns are also extracted, and they represent combinations of terms and types that represent opinions and relationships, such as positive comments about an organization.
- **Refine** the extracted concepts and fine-tune your extractions by directly manipulating one or more libraries containing word types, terms, synonyms, exclude lists, and other linguistic constructs. As mentioned earlier, text analysis is an iterative process where refining your libraries and dictionaries directly produces results that are fine-tuned to your data.
- **Categorize** your responses by creating and editing categories manually using category rules, code frames, and/or automatically using category building techniques. The categories represent higher-level concepts that capture the chief ideas and attitudes expressed by the respondents.
- **Export** your categories along with the ID variable into common file formats for further analysis and graphing in other applications. The output can be a set of multiple-response variables, as either an SPSS Statistics or Microsoft Excel file.

The Text Analysis Window

The application interface is made up of two windows. The first is the text analysis window, where you will perform the bulk of your work. In this window, you can analyze each question in your data. For each question, you can extract concepts, types, and patterns, and then categorize your responses.

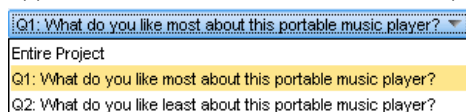
When you start the application, you are presented with a screen in which you can open an existing project or create a new one. If you choose to create a new project, a wizard opens to guide you through the project creation process. For more information, see the topic “Creating Projects” in Chapter 3 on p. 25.

Figure 2-1
Text analysis window at product launch



Once you have imported data, you can look at: the Question view(s) or the Entire Project view. You can change views by selecting one from the drop-down list on the toolbar in the text analysis window or by selecting the view from the View menu. The text that appears in the list box is taken from the variable label for each question.

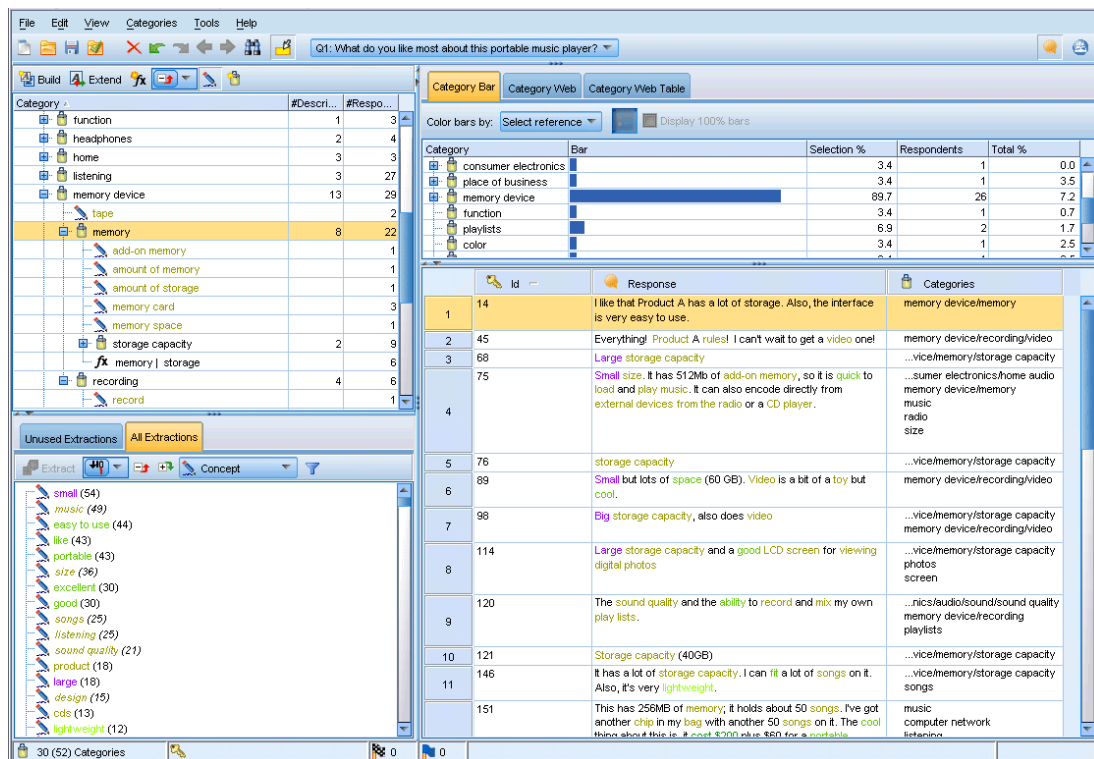
Figure 2-2
Application toolbar with view selector drop-down list



The Question View

The Question view provides you with a space in which you can analyze and categorize the responses to a particular question. After creating a new project, the Question view appears with the responses for the first open-ended text variable that you imported. You can select this view from the drop-down list or choose a question name from the View > Question > menu.

Figure 2-3
Question view



The operations that you perform in the Question view revolve around three elements: extraction results, categories, and response data. In order to help you work and analyze each of the elements independently, this window is divided into four panes:

Categories Pane

Located in the upper left corner, this pane presents an area in which you can create and manage any categories you build. After extracting the concepts, types, and patterns from your text data, you can begin building categories by using automatic techniques, such as semantic networks and concept inclusion, or by creating them manually. You can click and expand a category to see all of the descriptors that make up its definition, such as concepts, types, and rules. When you select a category or descriptor, you can then display information about corresponding records in the Data and Visualization panes. For more information, see the topic “The Categories Pane” in Chapter 6 on p. 92.

Figure 2-4
Categories pane: Expanded category definition

Category	#Descri...	#Respo...
function	1	3
headphones	2	4
home	3	3
listening	3	27
memory device	13	29
tape		2
memory	8	22
add-on memory		1
amount of memory		1
amount of storage		1
memory card		3
memory space		1
storage capacity	2	9
memory storage		6
recording	4	6
record		1

Extracted Results Pane

Located in the lower left corner, this area presents the extraction results. When you run an extraction, the extraction engine reads through the text data, identifies the relevant concepts, and assigns a type to each. **Concepts** are words or phrases extracted from your text data. **Types** are semantic groupings of concepts stored in the form of type dictionaries. When the extraction is complete, concepts, types, and patterns appear in the Extraction Results pane. Concepts and types are color-coded to help you identify what type they belong to. You can use these concepts, types, and patterns you collect here to build your categories. For more information, see the topic “Extracted Results: Concepts, Types, and Patterns” in Chapter 5 on p. 77.

Text mining is an iterative process in which extraction results are reviewed according to the context of the text data, fine-tuned to produce new results, and then reevaluated. Extraction results can be refined by modifying the linguistic resources. This fine-tuning can be done in part directly from the Extracted Results or Data pane but also directly in the Resource Editor view. For more information, see the topic “The Resource Editor Window” on p. 14.

Data Pane

Located in the lower right side of this view, it presents, in a tabular format, the response data for the selected open-ended question. By default, the Data pane shows three columns (record IDs, text responses, and assigned categories). The number of responses that appear in this pane are filtered according to what you have selected in another pane. While you can view the data that you imported in this pane, you cannot edit, delete, or append to the records. For more information, see the topic “The Data Pane” in Chapter 6 on p. 95.

Visualization Pane

Located in the upper right side of this view, it is hidden by default. You can display this pane (choose View > Visualization). This pane offers three unique views of how responses fit into categories or how categories may share responses (web chart, bar chart, and table) according to the selections you make in the other panes. For more information, see the topic “Visualizing Graphs” in Chapter 7 on p. 159.

Depending on whether you chose the extraction option in the New Project wizard, you may or may not have extraction results in the lower left hand pane. Click Extract in the Extraction Results pane to begin extracting. After extracting, you can review the results to see if any fine-tuning is necessary, such as grouping synonyms under one concept name or excluding common, uninteresting concepts from the list.

Once you are satisfied with the extraction results, you can begin categorizing your responses manually by dragging and dropping concepts as categories or using category building techniques, such as concept inclusion and a semantic network.

The Entire Project View

The Entire Project view provides an overview of all of the variables that you imported to the project. You can select this view from the drop-down list or from the View menu (View > Entire Project). In this view, you can review the data you imported, change a variable’s role (for example, from question to reference variable), and assign labels to the variables. For more information, see the topic “Viewing Project Data” in Chapter 4 on p. 49. While you can view the data that you imported in this view, you cannot edit, correct, delete, or append to the records.

Note: To view the entire contents of a cell in this view, you can hover the mouse over the cell. A ToolTip displays the cell contents.

Figure 2-5
Entire Project view

	Respondent ID	Q1: What do you like most about this portable music player?	Q2: What do you like least about this portable music player?	REF1: Product	REF2: Age
1	1	little, light	expensive	Other	25-34
2	2	The battery power is great.	The screen is hard to see when outside.	Product E	35-44
3	3	cost and size	difficult software	Other	25-34
4	4	Having all my CDs in the palm of my hand!	Nothing, I love it!	Product A	35-44
5	5	The shuffle mode.	Battery life seems shorter than advertised.	Product A	35-44
6	6	Battery life. Portability. Accessories. Style.	Ubiquitousness; everyone has one.	Product A	25-34
7	7	I like its ability to store all of my music. I also...	I wish the 40GB model was still available. I ...	Product A	35-44
8	8	portability, capacity, sound quality, durability	it doesn't have a light.	Other	35-44
9	9	Small, great sound, capacity.	Nothing, I love it.	Product A	25-34
10	10	Able to hold all of my songs in one place.	It is in the shop due to a hardware failure.	Product A	35-44
11	11	It's portable! I can take it anywhere.	smudges on the display	Product A	45-54
12	12	Living in my own little world	Battery life	Product A	35-44
13	13	mobility	Technical difficulties setting it up initially an...	Product A	35-44
14	14	I like that Product A has a lot of storage. Al...	It is a little heavy, and the battery life isn't lo...	Product A	25-34
15	15	It holds a ton of music.	Battery life.	Product A	25-34
16	16	It's fun to use	nothing	Product A	45-54
17	17	its cool	battery	Product A	35-44
18	18	lots of disk space	it was very expensive	Product A	25-34
19	19	Others think it is cool and it sounds great.	I find the controls hard to use.	Product B	<18
20	20	lightweight	so small afraid I'll lose it easily	Product A	45-54
21	21	easy to use	size	Product D	25-34
22	22	great accessories	high price	Product A	45-54
23	23	i can listen to my music wherever i want. i ...	i can't change the color of the outside. I wa...	Product A	25-34
24	24	supports standard for lossless compressio...	window scratches easily	Other	>54
25	25	very small and holds lots of songs	didn't come with a belt clip	Product C	18-24
26	26	its great. i can share music with my friends...	it is old now and big compared to the ones ...	Product A	18-24
27	27	I can listen to the old Ludwig Van without a...	it is often mistaken for a cell phone.	Product A	25-34
28	28	It offers lots of disk space for all of my CDs.	It is expensive.	Product A	18-24
29	29	Always having a good collection of music a...	Battery life is limited	Product A	25-34
30	30	It's portable and the device is well-designed.	You have to shut down the device by holdi...	Product A	25-34
31	31	Its portability enables me to listen to my mus...	I think that it could be lighter and less expen...	Product B	>54

The Resource Editor Window

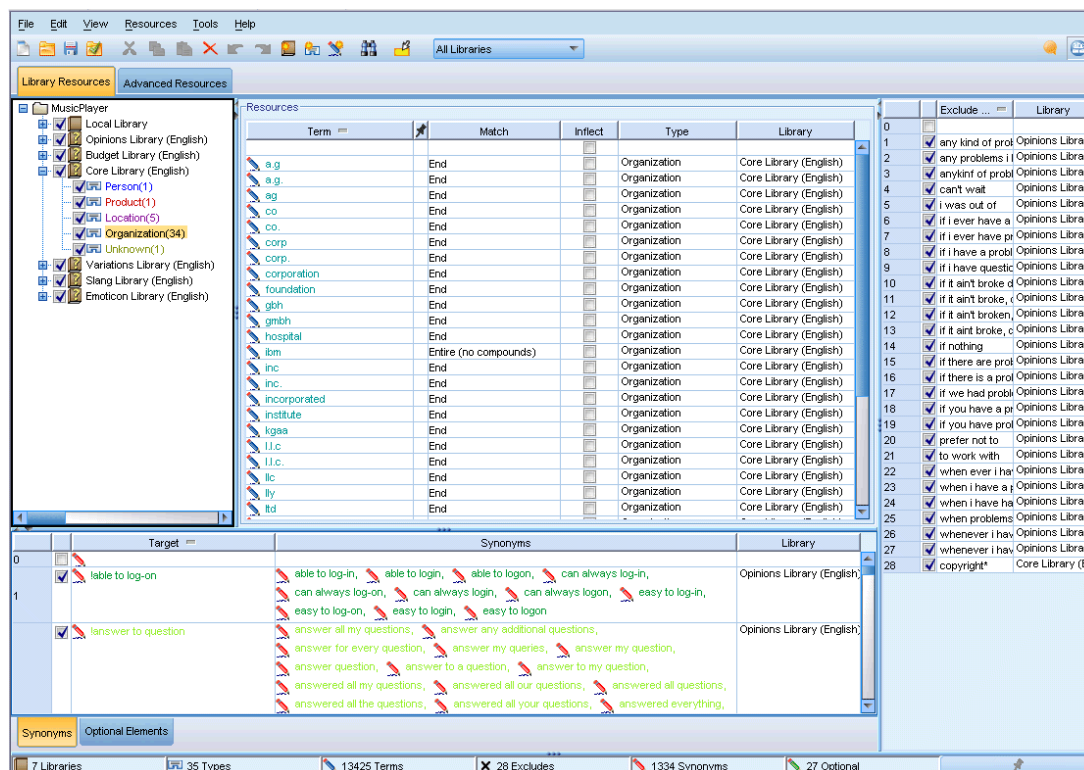
IBM® SPSS® Text Analytics for Surveys rapidly and accurately captures key concepts from text data using a robust extraction engine. This engine relies heavily on linguistic resources to dictate how large amounts of unstructured, textual data should be analyzed and interpreted.

The Resource Editor view is where you can view and fine-tune the linguistic resources used to extract concepts, group them under types, discover patterns in the text data, and much more. IBM® SPSS® Text Analytics for Surveys offers several preconfigured resource templates. Also, in some languages, you can also use the resources in a text analysis packages. For more information, see the topic “Using Text Analysis Packages” in Chapter 3 on p. 40.

Since these resources may not always be perfectly adapted to the context of your data, you can create, edit, and manage your own resources for a particular context or domain in the Resource Editor. For more information, see the topic “Working with Libraries” in Chapter 9 on p. 195. Some users may use this window infrequently, since the resources delivered with the product often suffice. Furthermore, much of the simple library work that you may perform can be done directly from the Extraction Results pane in the text analysis window.

To simplify the process of fine-tuning your linguistic resources, you can perform common dictionary tasks directly from the Text Analytics view through context menus in the Extraction Results and Data panes. For more information, see the topic “Refining Extraction Results” in Chapter 5 on p. 84.

Figure 2-6
Resource Editor view



The operations that you perform in the Resource Editor view revolve around the management and fine-tuning of the linguistic resources. These resources are stored in the form of templates and libraries. The Resource Editor view is organized into four parts: Library Tree pane, Type Dictionary pane, Substitution Dictionary pane, and Exclude Dictionary pane.

The interface is organized into four parts, as follows:

1. Library Tree pane. Located in the upper left corner, this pane displays a tree of the libraries. You can enable and disable libraries in this tree as well as filter the views in the other panes by selecting a library in the tree. You can perform many operations in this tree using the context menus. If you expand a library in the tree, you can see the set of types it contains. You can also filter this list through the View menu if you want to focus on a particular library only.

2. Term Lists from Type Dictionaries pane. Located to the right of the library tree, this pane displays the term lists of the type dictionaries for the libraries selected in the tree. A **type dictionary** is a collection of terms to be grouped under one label, or type, name. When the extraction engine reads your text data, it compares words found in the text to the terms in the type dictionaries. If an extracted concept appears as a term in a type dictionary, then that type name is assigned. You can think of the type dictionary as a distinct dictionary of terms that have something in common. For example, the <Location> type in the Core library contains concepts such as *new orleans*, *great britain*, and *new york*. These terms all represent geographical locations. A library can contain one or more type dictionaries. For more information, see the topic “Type Dictionaries” in Chapter 10 on p. 207.

3. Exclude Dictionary pane. Located on the right side, this pane displays the collection of terms that will be excluded from the final extraction results. The terms appearing in this exclude dictionary do not appear in the Extraction Results pane. Excluded terms can be stored in the library of your choosing. However, the Exclude Dictionary pane displays all of the excluded terms for all libraries visible in the library tree. For more information, see the topic “Exclude Dictionaries” in Chapter 10 on p. 222.

4. Substitution Dictionary pane. Located in the lower left, this pane displays synonyms and optional elements, each in their own tab. Synonyms and optional elements help group similar terms under one lead, or target, concept in the final extraction results. This dictionary can contain known synonyms and user-defined synonyms and elements, as well as common misspellings paired with the correct spelling. Synonym definitions and optional elements can be stored in the library of your choosing. However, the substitution dictionary pane displays all of the contents for all libraries visible in the library tree. While this pane displays all synonyms or optional elements from all libraries, The substitutions for all of the libraries in the tree are shown together in this pane. A library can contain only one substitution dictionary. For more information, see the topic “Substitution/Synonym Dictionaries” in Chapter 10 on p. 217.

Note:

- If you want to filter so that you see only the information pertaining to a single library, you can change the library view using the drop-down list on the toolbar. It contains a top-level entry called All Libraries as well as an additional entry for each individual library. For more information, see the topic “Viewing Libraries” in Chapter 9 on p. 199.

Setting Options

You can set general options for IBM® SPSS® Text Analytics for Surveys in the Options dialog box. This dialog box contains the following tabs:

System tab contains options for default library lists, autosaving, saving extraction results, delimiters, and the interface language.

Display tab contains options for the colors used in the interface.

Sounds tab contains options for sound cues.

Translation tab contains options for translation connections.

To Edit Options

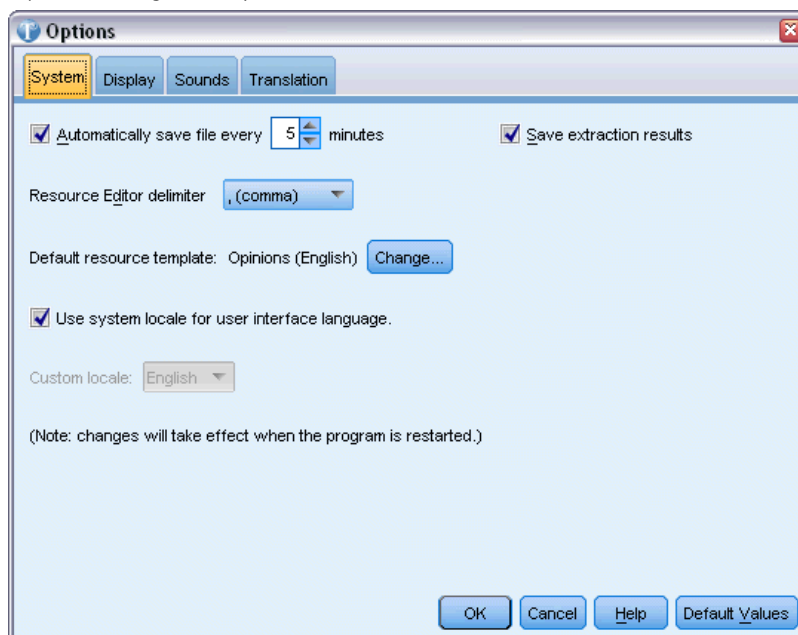
- ▶ From the menus, choose Tools > Options. The Options dialog box opens.
- ▶ Select the tab containing the information that you want to change.
- ▶ Change any of the options.
- ▶ Click OK to save the changes.

Options: System Tab

On this tab, you can define many project settings, including:

- Adding or removing libraries that should appear in all new projects by default
- Enabling or disabling the autosave recovery feature
- Enabling or disabling the saving of extraction results
- Defining the global delimiter that will be used in the Resource Editor to separate elements

Figure 2-7
Options dialog box: System tab



Automatically save file every n minutes. Select this option to have the product automatically create a temporary saved version of the open project file in case of machine failure. Also, set the number of minutes between each save. If you enable this feature and the product closes unexpectedly or you experience a machine issue, the next time you launch the product, you will be given a chance to open and work with a recovery version of your file.

Save extraction results. Select this option to save the results of your extractions in your project. This can save time when you are still working on your categories. However, it can add time when you are loading, and it can increase the size of your project. As a security measure, these extraction results are encrypted during the save process and placed in the database. This procedure makes it difficult for someone, even an advanced user, to come across any data in the database. Furthermore, extraction results are never presented in IBM® SPSS® Text Analytics for Surveys until the data source has been located by the application. Therefore, if the data are password-protected, a user must enter the user name and password for this data source before the extraction results appear on the screen.

- *Saving* is most advantageous for time-efficiency. Given that the extraction process can take a while to complete when working with larger data sets, saving provides you with immediate access to the results whenever you reopen your project. However, you may notice a slightly longer wait time when opening a project.
- *Not saving* is used whenever you do not want any of the response text to reside anywhere other than in the original data file, even though security measures are in place.

Resource Editor Delimiter. Select the character to be used as a delimiter when entering elements, such as terms, synonyms, and optional elements, in the Resource Editor.

Resource template. If you did not select a text analysis package, a set of default resources will be used. These resources are stored in a template. Click Change to select a different default resource template. Then, in Change Templates dialog, select the line in the table for the template you want to use and click OK.

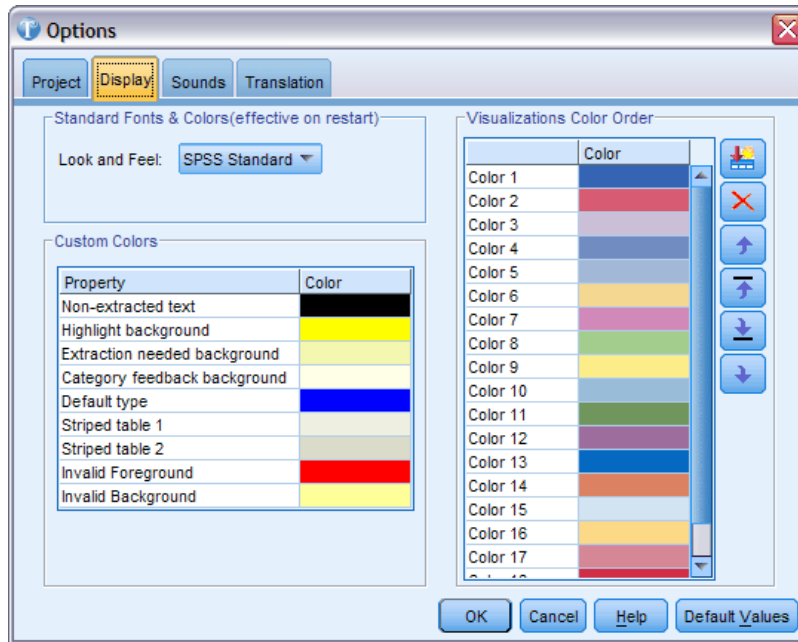
Use system locale for user interface language. Select this option to have SPSS Text Analytics for Surveys use your system's locale details to provide the language used on the interface. Alternatively, you can deselect this option and then choose a different interface language. For example, you may do this if you are analyzing information recorded in a different language from your system locale and want to run SPSS Text Analytics for Surveys in that language.

Note that changes made to this option will not take effect until you restart SPSS Text Analytics for Surveys.

Options: Display Tab

On this second tab, you can edit options affecting the overall look and feel of the application and the colors used to distinguish elements.

Figure 2-8
Options dialog box: Display tab



Standard Fonts & Colors (effective on restart). Options in this control box are used to specify the color scheme and look displayed. Options selected here do not take effect until you close and restart the application. Choose from:

- **SPSS Standard (default)**, a design common across SPSS brand (a part of IBM Corp.) products.
- **SPSS Classic**, a design familiar to users of earlier versions of this application.
- **Microsoft Windows**, a Microsoft Windows design that may be useful for increased contrast in the stream canvas and palettes.

Custom Colors. Edit the colors for elements appearing onscreen. For each of the elements in the table, you can change the color. To specify a custom color, click the color area to the right of the element you want to change and choose a color from the drop-down color list.

- **Non-extracted text.** Response text that was not extracted yet visible in the Data pane.
- **Highlight background.** Text selection background color when selecting elements in the panes or text in the Data pane.
- **Extraction needed background.** Background color of the Extraction Results pane indicating that changes have been made to the libraries and an extraction is needed.
- **Category feedback background.** Category background color that appears after an operation, such as dragging and dropping responses and forcing responses from the Data pane into the Categories pane.
- **Default type.** Default color for types and terms appearing in the Data pane and Extraction Results pane. This color appears in the interface whenever the Unknown type or any of the associated concepts appear. This color will also apply to any custom types that you create in the Resource Editor. You can override this default color for your custom type dictionaries by

editing the properties for these type dictionaries in Resource Editor. For more information, see the topic “Creating Types” in Chapter 10 on p. 209.

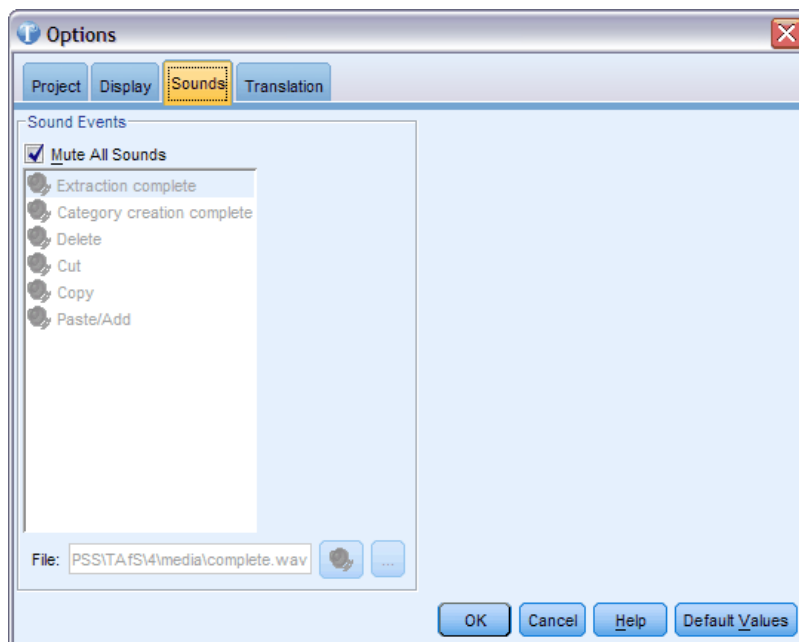
- **Striped table 1.** First of the two colors used in an alternating manner in the table in the Edit Forced Terms dialog box in order to differentiate each line.
- **Striped table 2.** Second of the two colors used in an alternating manner in the table in the Edit Forced Terms dialog box in order to differentiate each line.
- **Invalid foreground.** Color for the text of duplicate entries in the Code Frame Manager, indicating an error.
- **Invalid background.** Color for the background highlighting of duplicate entries in the Code Frame Manager, indicating an error.

Visualizations Color Order. If you use the category bar chart in the visualization pane and also select a reference variable, you can see each of the possible values for the reference variable in a legend at the bottom of the pane. Each value is also color coded to help you visually distinguish it in the bar chart. You can change these default colors here. For more information, see the topic “Visualizing Graphs” in Chapter 7 on p. 159.

Options: Sounds Tab

On this tab, you can edit the sounds used in the product. Under Sound Events, you can specify a sound to be used to notify you when an event occurs. By turning sounds on or off or assigning specific sounds, you can control the way you are alerted to particular operations in the software. For example, you can activate sounds for events such the end of the extraction process, the end of an automatic categorization technique, or more common tasks, such as cut, paste, copy, and delete.

Figure 2-9
Options dialog box: Sounds tab



A number of sounds are available. Use the ellipsis button (...) to browse for and select a sound. The .wav files used to create sounds for IBM® SPSS® Text Analytics for Surveys are stored in the /media subdirectory of the installation directory. If you do not want sounds to be played, select Mute All Sounds. Sounds are muted by default.

Options: Translation Tab

Important! Translation is only available into English.

On this tab, you can define and manage the Language Weaver translation server connection that you can reuse anytime you translate. Once you define a connection here, you can quickly choose a language pair connection at translation time without having to reenter all of the connection settings.

A language pair connection identifies the source and translation languages as well as the URL connection details to the server. For example, *Chinese - English* means that the source text is in Chinese and the resulting translation will be in English. You have to manually define the connection for the Language Weaver server you access through the Language Weaver online services.

The translation results are stored in the directory location defined in this dialog. You can manage your translation files directly in that directory and/or specify a different directory here.

Figure 2-10
Options dialog box: Translation tab

The screenshot shows the 'Options' dialog box with the 'Translation' tab selected. The dialog contains the following fields and buttons:

- Connection URL:**
- User ID:**
- API Key:**
- Translation directory:**

At the bottom of the dialog are buttons for **OK**, **Cancel**, **Help**, and **Default Values**.

Connection URL. Enter the Server URL or web address to the Language Weaver online server.

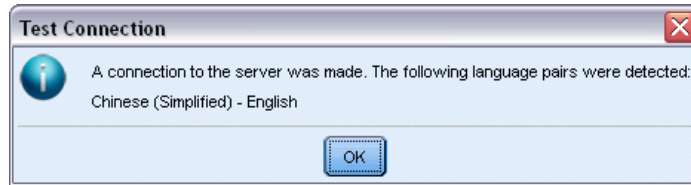
User ID. Enter the unique ID provided to you by Language Weaver.

API Key. Enter the key provided to you by Language Weaver.

Test. Click Test to verify that the connection is properly configured and to see the language pair(s) that are found on that connection.

Translation directory. Click Browse to change to a different directory or type the folder path directly into the field.

Figure 2-11
Successful connection message



Microsoft Internet Explorer Settings for Help

Microsoft Internet Explorer Settings

Most Help features in this application use technology based on Microsoft Internet Explorer. Some versions of Internet Explorer (including the version provided with Microsoft Windows XP, Service Pack 2) will by default block what it considers to be “active content” in Internet Explorer windows on your local computer. This default setting may result in some blocked content in Help features. To see all Help content, you can change the default behavior of Internet Explorer.

- ▶ From the Internet Explorer menus choose:
Tools > Internet Options...
- ▶ Click the Advanced tab.
- ▶ Scroll down to the Security section.
- ▶ Select (check) Allow active content to run in files on My Computer.

Part II:
Text Analysis

Creating Projects and Packages

Creating Projects

In IBM® SPSS® Text Analytics for Surveys, you will work with and categorize survey data. To do so, you will create projects into which you will import the data from your data source, select some variables, and choose categories and resources. Once you create a project, you can fine tune your resources and categories until you feel you have the final set of categories. A project can contain the following elements: survey data, linguistic resources, extraction results, and categories.

Survey Data

The imported survey data source is referenced within the project, but the survey data are not stored in the project. Instead, when a project is opened, survey data are reread from the original data source. While the most important variables in this context are the open-ended text variables, the unique ID variable is retained, as are any reference variables (such as demographic variables) specified when the data were imported. Values for all of these variables can be displayed within the Data pane in the Question view in the text analysis window or in the Entire Project view in the text analysis window.

Linguistic Resources

Proprietary and user-customized libraries containing lists of terms, synonyms, lists of excluded words, and type declarations are stored in a project and can be modified. In addition, certain compiled resources are used to process text and are also stored in the project and cannot be edited. Libraries can be published, which makes them publicly available within the database for use within other projects. A published library can be added to other projects.

Extracted Results

The extraction results are key words and phrases (concepts), their semantic groupings (types), and their relationships (patterns) that are identified and extracted from the text responses. These extraction results are part of the project and are the basis of category creation. By default, extraction results are saved in the project, but if you think they make the project file size too large, you can turn off this saving feature and reextract the next time you open the project. For more information, see the topic “Saving Extraction Results” in Chapter 5 on p. 84.

Categories

Text responses are placed into categories that can be created automatically by using category building techniques, manually through drag-and-drop operations, by importing category definition files, or by using the Code Frame Manager. If you choose not to save the extraction results, whenever a project is reopened, the category definitions will remain but response counts for any

parts of the definitions that came from the extraction results will be displayed with a question mark (?) until you reextract.

Preparing Your Data

Before importing your data into IBM® SPSS® Text Analytics for Surveys, please review the following considerations:

- **Input data.** In order to import your data source into SPSS Text Analytics for Surveys, it must contain certain basic elements, such as an ID variable and at least one open-ended question. The ID variable must contain only unique values. Any duplicates will cause the import to fail. You can import multiple open-ended questions as well as reference variables. For more information, see the topic “Selecting Variables” on p. 32.
- **Spelling errors.** While the program accommodates some spelling errors, we recommend that you correct such errors before importing your data into the program. Spelling errors can cause problems in text analysis for humans as well as for software programs. The more spelling errors you can correct beforehand, the more reliable the resulting categories are. You can also create synonyms with the correct spelling of a word and the commonly misspelled variations in the program. In fact, many common misspellings are predefined in the Core library. If you are unsure of how much effort to expend on spell checking, you can run some experiments with a smaller sample of responses to see how much the analysis is affected by spelling errors.
- **Blank responses.** It is not uncommon to find blank responses in open-ended survey data. Although blank responses provide no information, they can still be useful. For example, you might find it interesting to know how many people did not respond to a question or what type of people did not respond. However, since SPSS Text Analytics for Surveys uses text to extract terms and categorize responses, these blank responses cannot easily be categorized.

One approach is to replace all of the blank responses with the word *blank* or some other suitable term in your data before importing. Then, after the data are imported, you can create a new type that represents a blank response, with the word *blank* (or whatever word you inserted) being the term that represents that type.

Another option involves forcing blank responses into a category. After you categorize the responses, blank responses will initially be uncategorized. You can create a new Blank category manually by right-clicking in the Categories pane. Then, after selecting all of the blank responses, you can force them all into the new Blank category.

- **Multiple-response questions.** While open-ended questions usually stand on their own, this is not always the case. Sometimes open-ended questions are used as a multiple-response set. For example, if you ask a respondent to “Tell us three things we can improve about this hotel” and provide three separate spaces (variables) in which to reply, this represents a multiple-response question.

Since SPSS Text Analytics for Surveys analyzes each question variable separately, you could reuse the categories and linguistic resources created to analyze the first response to categorize the second and third responses. However, this may not be the most efficient method. You may want to consider combining all three response variables into one variable before importing the data into the program. If you combine them, please verify that you have at least a space between the last word in one response and the first word in the next, or preferably a period.

Since this may be a time-consuming task with larger data sets, consider combining the responses as the data file is being created rather than afterward.

- **Response samples.** The greater the number of responses and the longer each is on the average, the more time an extraction or categorization will require. To work more efficiently, when your file size is large (perhaps 1,500 cases or more), you can take a random sample first and use that smaller subset of responses to do a first pass at the analysis.

A smaller sample is often perfectly adequate to decide how to edit the linguistic resources. And once you have categorized on the smaller data file, you can read in the full file and reextract, which will automatically categorize many of the responses. Then you can look for responses that do not fit the categories you had created and make any needed adjustments. The size of the random sample can vary, but 300 or so cases will usually be adequate.

Important! There are other considerations regarding the text analysis process as a whole. For more information, see the topic “Preparing for Text Analysis” in Chapter 1 on p. 7.

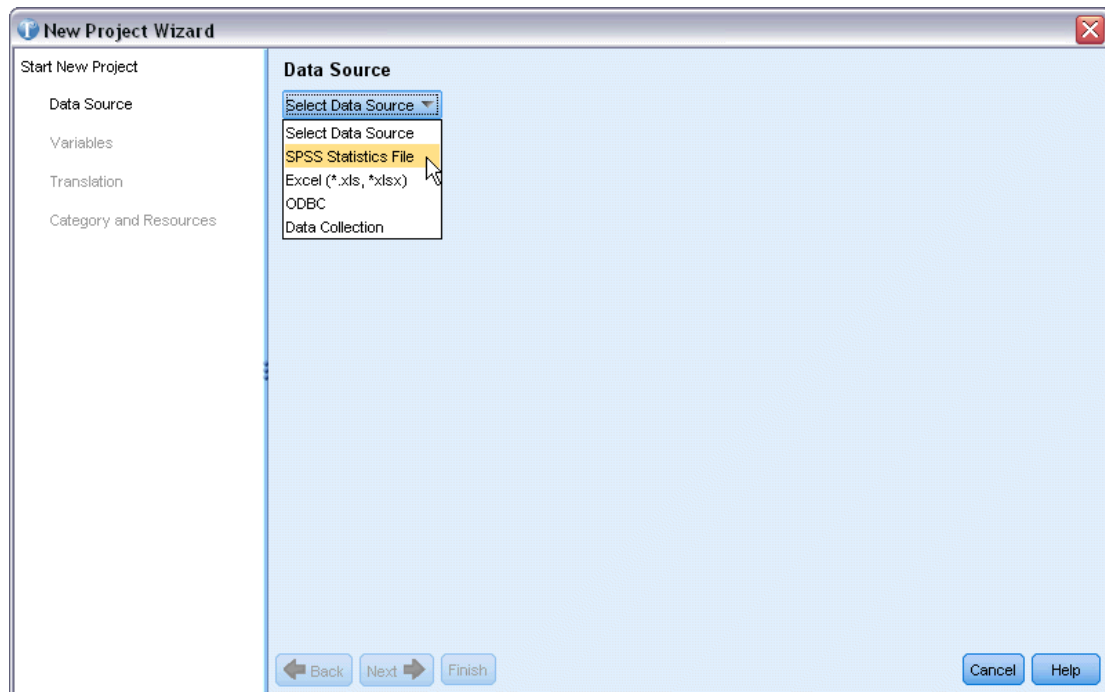
Starting New Projects

In order to begin categorizing your response data, you must first create a project. A wizard guides you through data source and variable selection, category and resource specifications, and more. Before you begin creating your project, you may want to prepare your data.

To Start a New Project

- ▶ From the menus choose File > New Project. Alternately, click Start a New Project from the startup screen if no projects are open. The New Project Wizard appears.

Figure 3-1
New Project Wizard



- ▶ Begin by selecting the data source type from the Select Data Source drop-down list. For more information, see the topic “Selecting Data Sources” on p. 28.

Selecting Data Sources

When the wizard opens, you begin by selecting a data source. IBM® SPSS® Text Analytics for Surveys was optimized to process data sets of up to 10,000 records, although performance will vary based on the volume of text contained in these records. See the installation instructions for performance statistics and recommendations.

Important! An ID variable with a unique value for each record must be present in order to import the data.

You can choose one of the following data sources:

- **SPSS Statistics files** (*.sav).
- **Microsoft Excel files** (*.xls / *.xlsx).
- **ODBC** (Microsoft Open Database Connectivity protocol) database.
- **Data Collection** data model. This option is available only if you have the data model installed.

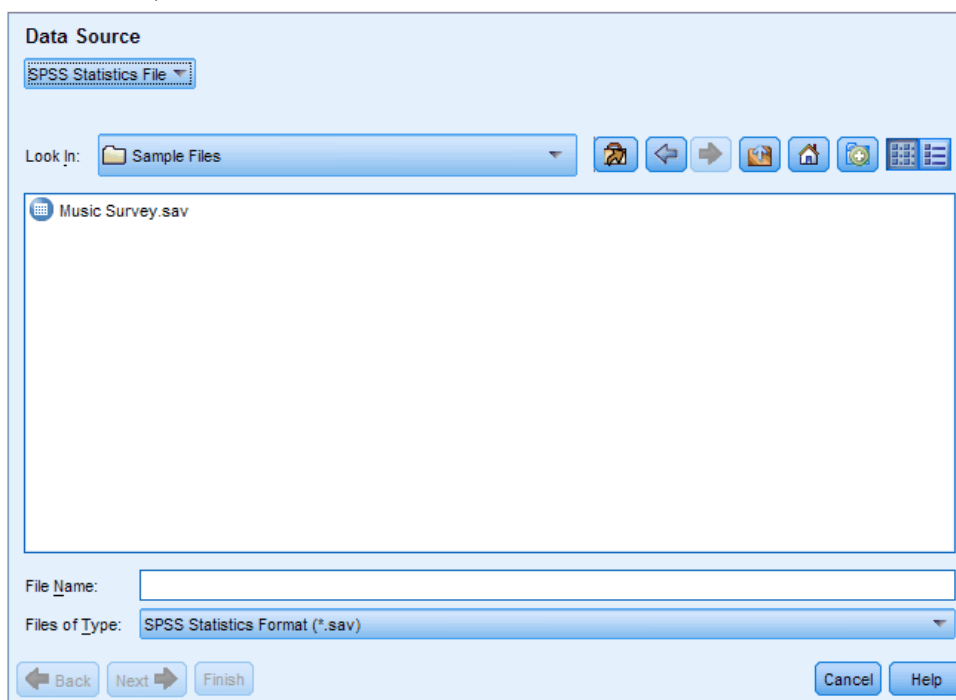
Using IBM SPSS Statistics Files

You can import an IBM® SPSS® Statistics (.sav) file into IBM® SPSS® Text Analytics for Surveys. An ID variable with a unique value for each record must be present in order to import the data.

Important! You cannot import SPSS Statistics (.sav) file with records exceeding 4000 characters.

Note: SPSS Text Analytics for Surveys was optimized to process data sets of up to 10,000 records, although performance will vary based on the volume of text contained in these records. See the installation instructions for performance statistics and recommendations.

Figure 3-2
Data source options for IBM SPSS Statistics files



To Get Data from IBM SPSS Statistics

- ▶ In the first screen of the wizard, select SPSS Statistics file from the drop-down list. The wizard displays the options for SPSS Statistics files.
- ▶ From the Look In drop-down list, select the drive and folder in which the file is located.
- ▶ Select the file from the list. It will appear in the File Name text box.
- ▶ Click Next to select variables. For more information, see the topic “Selecting Variables” on p. 32.

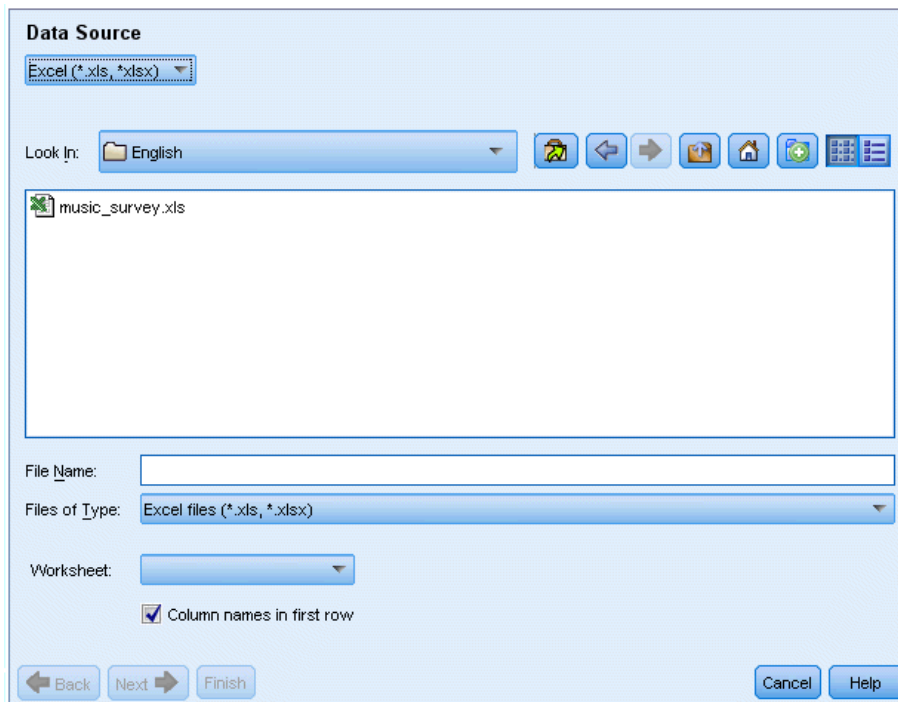
Using Microsoft Excel Files

You can import a Microsoft Excel (.xls / .xlsx) file into IBM® SPSS® Text Analytics for Surveys. An ID variable with a unique value for each record must be present in order to import the data.

Important! During the Microsoft Excel file import, you can select an option for Column Names in First Row. To use this option, the very first row of the worksheet must contain column names—not the row just above where the data begin. For example, if your data and column names begin on row 7, you must delete rows 1–6 before importing in order to import the file correctly.

Note: SPSS Text Analytics for Surveys was optimized to process data sets of up to 10,000 records, although performance will vary based on the volume of text contained in these records. See the installation instructions for performance statistics and recommendations.

Figure 3-3
Data source options for Microsoft Excel files



To Get Data from Microsoft Excel

- ▶ In the first screen of the wizard, select Excel from the drop-down list. The wizard displays the options for Microsoft Excel files.
- ▶ From the Look In drop-down list, select the drive and folder in which the file is located.
- ▶ Select the file from the list. It will appear in the File Name text box.
- ▶ Select the worksheet from the drop-down list. You can only import data from a single worksheet. To work with data on multiple worksheets, you must create multiple projects.

- ▶ If the first row of this worksheet contains the column headers, select Column Name in First Row. To use this option, the very first row of the worksheet must contain column names—not the row just above where the data begin. For example, if your data and column names begin on row 7, you must delete rows 1–6 before importing in order to import the file correctly. The application can use these (or a converted version if the column headings do not conform to IBM® SPSS® Statistics variable-naming conventions) as the variable names. If not, the application will use the spreadsheet column letters as identifiers.
- ▶ Click Next to select variables. For more information, see the topic “Selecting Variables” on p. 32.

Using Data through ODBC

Data from database sources, commonly databases, are easily imported into IBM® SPSS® Text Analytics for Surveys. Any database that uses Open Database Connectivity (ODBC) drivers can be read directly by the product after the proper drivers are installed on the machine on which SPSS Text Analytics for Surveys is installed. An ID variable with a unique value for each record must be present in order to import the data.

Note: SPSS Text Analytics for Surveys was optimized to process data sets of up to 10,000 records, although performance will vary based on the volume of text contained in these records. See the installation instructions for performance statistics and recommendations.

Figure 3-4
Data source options for ODBC

Data Source

ODBC

Source(DSN)

Name	Description
Visio Database Samples	Microsoft Access Driver (*.MDB)
MS Access Database	Microsoft Access Driver (*.mdb)
Excel Files	Microsoft Excel Driver (*.xls)
dBASE Files	Microsoft dBase Driver (*.dbf)

ODBC

User:

Password:

Table

SQL

Back Next Finish Cancel Help

To Use Via ODBC

- ▶ In the first screen of the wizard, select ODBC from the drop-down list. The wizard displays the options for ODBC.
- ▶ Specify the data source by selecting it from the list of registered ODBC sources or by typing the name into the Source (DSN) text box. If you need to register new data sources that do not appear in the list, click ODBC. This will open the ODBC Data Source Administrator, which is found on most Microsoft Windows computers. If it is not found, you cannot use the ODBC import. Consult the Microsoft Windows Help system for more information.
- ▶ If the data source is password protected, you must enter a user name and password. You will be required to do so each time you open the project, since, for security reasons, the user name and password are not stored in the project.
- ▶ Select your data in one of two ways: Table or SQL. You can select a table directly or use SQL commands to select data.
- ▶ Click Next to select variables. For more information, see the topic “Selecting Variables” on p. 32.

Using IBM SPSS Data Collection Data**To Import Via IBM SPSS Data Collection**

- ▶ In the first screen of the wizard, select Data Collection from the drop-down list. The IBM® SPSS® Data Collection data model option is available only if you have the data model installed with another product.

Selecting Variables

After selecting the data source, the next step is to specify the variables to be imported. Three types of variables can be imported into a project.

Unique ID Variable (Required)

The ID variable is a unique numeric or string key that identifies each respondent. The data file does not need to be ordered by the unique ID variable to successfully read it. After being read into the program, the records can be sorted by various criteria. For more information, see the topic “Sorting Variables” in Chapter 4 on p. 50. This ID variable is required to import data. Each imported record (or case) must have a unique ID value.

Two situations will cause the import to fail:

- Duplicate ID values detected
- Records with blank ID values

Note: If a duplicate ID is detected and you have IBM® SPSS® Statistics installed on your computer, you can use the Identify Duplicate Cases procedure in that product to identify duplicates and then use the options to indicate which records should be retained (primary cases).

Open-Ended Text Variable(s) (Required)

The open-ended text variables represent the text responses to the question(s) in the survey. At least one of these variables is required to import data. These variables can be string or long-string variables in SPSS Statistics, columns containing general or text cells in Microsoft Excel, or text or note fields from databases. Each open-ended text variable will be analyzed separately. There is a 4,000-character limit on the size (width) of each text variable imported from a .SAV file.

Reference Variable(s) (Optional)

The reference variables are additional, optional variables, generally categorical, that can be imported for reference purposes. Reference variables are not used in text analysis but provide supplemental information describing the respondent, which may aid understanding and interpretation. Demographic variables are often included as reference variables, since they can contribute to understanding which terms or categories are being used by which groups of individuals. Examples are sex, department, occupation, and course of study (for student and training evaluations). You can view all of the reference variables after importing in the Entire Project view. You can also display reference variables in the Data pane of the Question view. Additionally, you can select reference variables in the bar chart in the visualization pane to be able to drill down to a subset of respondents.

Note: Reference variables read from an SPSS Statistics data file will have variable labels (if supplied) appearing as column headings and their value labels (if supplied) displaying in the cells of the Data pane.

Figure 3-5
Selecting variables

Variables

Select the variables for your survey analysis.

ID
 Q1leisurefactors
 Q2businessfactors
 Q3customerservice
 Q4carcomments
 R1samecompany
 R2samecar
 Gender

Unique ID

Open Ended Text

Reference

Switch variable names/labels

Automatically extract:

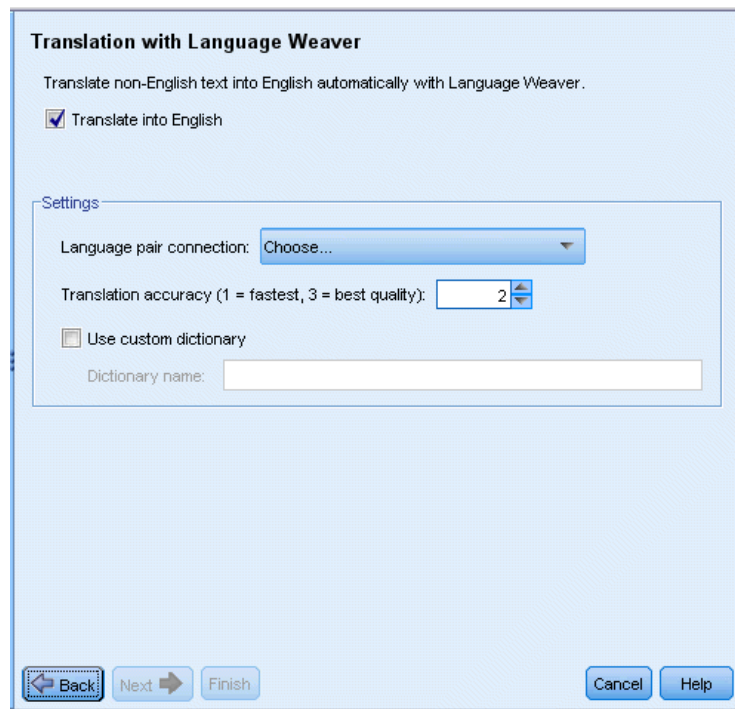
To Select Variables and Extraction Options

- ▶ From the list of available variables, select the variable that corresponds to the ID variable in your data set and click the arrow button to move it into the Unique ID box. The ID must be a unique number or alphanumeric string that distinguishes one record from another. If your data set contains duplicate IDs, an error message appears. In this case, you must clean your data before trying again.
- ▶ From the list of available variables, select one or more variables that correspond to the open-ended response variables and click the arrow button to move the variable(s) into the Open-Ended Text list. The variable(s) will each be imported as a separate question whose responses you will analyze and categorize.
- ▶ From the list of available variables, select one or more variables that correspond to the reference variables and click the arrow button to move the variable(s) into the Reference list. Reference variables are not used by the automated category building techniques. However, you can view their content and use them to help you make informed decisions when categorizing your responses.
- ▶ To view the variable labels instead of the variable names, click the button below the variable list on the left.
- ▶ To change the extraction setting, make a selection in the drop-down list. By default, First question only is selected, which means that if you have selected more than one open ended text variable, the extraction process will start automatically for the first question after the wizard ends. Extraction can take some time with larger data sets. Therefore, you may choose to extract None or All questions depending on the time available.
- ▶ Click Next > once you have selected all of your variables.

Translating into English

If you are working with non-English source text, you can connect to Language Weaver to translate into English. Translation is only available into English. You must have Language Weaver properly configured and with connections defined to translate. For more information, see the topic “Options: Translation Tab” in Chapter 2 on p. 21.

Figure 3-6
Translation options



To Translate Into English

- ▶ To translate the text data from a licensed language into English, select the Translate into English checkbox.
- ▶ From the Language Pair Connection list, select the connection for the Language Weaver language pair you want to use. If you have Language Weaver configured on your local machine, those language pairs will automatically appear in this list. You can add, change, or test the online services connection in the Translation tab of the Options dialog. For more information, see the topic “Options: Translation Tab” in Chapter 2 on p. 21.
- ▶ Specify the desired Translation accuracy. Choose a value of 1 to 3 indicating the level of speed versus accuracy you want. A lower value produces faster translation results but with diminished accuracy. A higher value produces results with greater accuracy but increased processing time. To optimize time, we recommend beginning with a lower level and increasing it only if you feel you need more accuracy after reviewing the results.
- ▶ If you have previously created custom dictionaries, held by Language Weaver, you can use them in connection with the translation. To choose a custom dictionary, select the Use custom dictionary checkbox and enter the Dictionary name. To use more than one dictionary, separate the names with a comma.
- ▶ In the New Project Wizard, click Next > to begin selecting categories and resources. For more information, see the topic “Selecting Categories and Resources” on p. 36.

- ▶ In the Change Data Set Wizard, click Finish to complete the data set change and to start the translation process.

To skip translation:

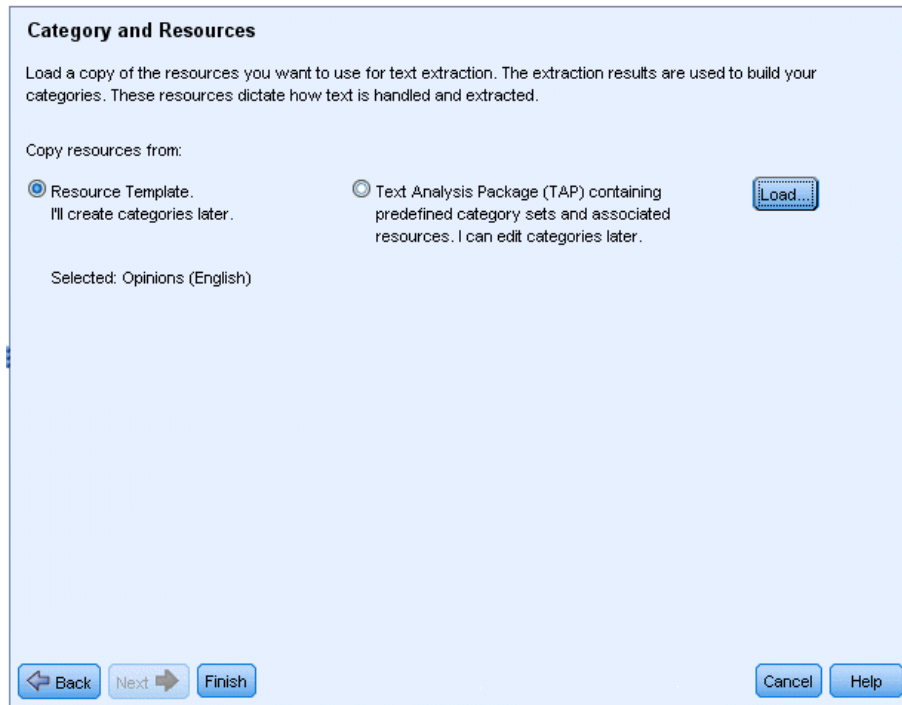
- ▶ Unselect the Translate into English option.
- ▶ In the New Project Wizard, click Next > to begin selecting categories and resources. For more information, see the topic “Selecting Categories and Resources” on p. 36.
- ▶ In the Change Data Set Wizard, click Finish to complete the data set change.

Selecting Categories and Resources

In this final step, you can select the linguistic resources that will be used to extract salient concepts and patterns from your text. Alternately, you can load a text analysis package (TAP), which not only includes the linguistic resources but also one or more predefined category sets that represent enhanced code frames. For more information, see the topic “Using Text Analysis Packages” on p. 40. Several prebuilt TAP files for English language text are offered by IBM® SPSS® Text Analytics for Surveys. Each TAP file shipped with this product is fine-tuned for a specific type of survey, such as employee, product, or customer satisfaction. You can also create your own TAPs for any text language supported by the product.

By default, a resource template is preloaded. You can change the default resource template that is proposed in the first tab of the Options dialog. For more information, see the topic “Setting Options” in Chapter 2 on p. 16. You can load a different resource template or choose a TAP instead.

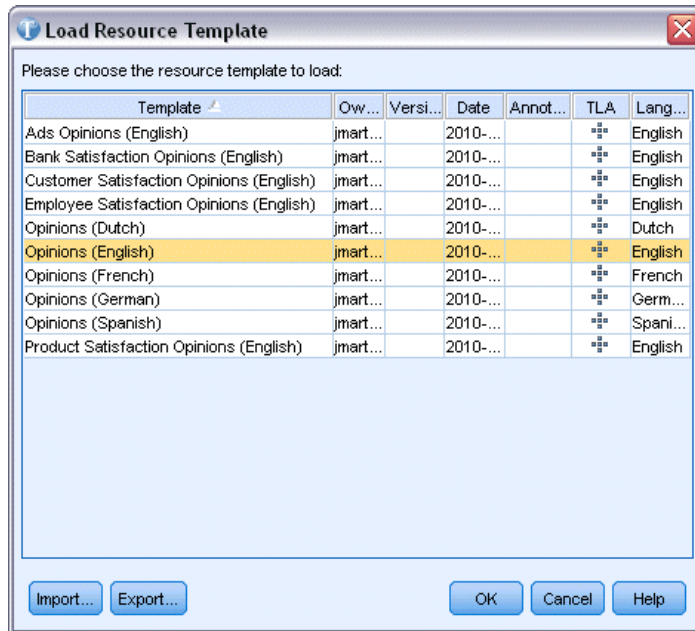
Figure 3-7
Selecting Resources



To select a different resource template:

- ▶ To load a different resource template, make sure the Resource Template option is selected and click Load. The Load Resource Template dialog opens.

Figure 3-8
Load resource template



- ▶ In the Load Resource Template dialog, select the template you want to use and click OK. The dialog closes and the wizard now shows the new template you selected.

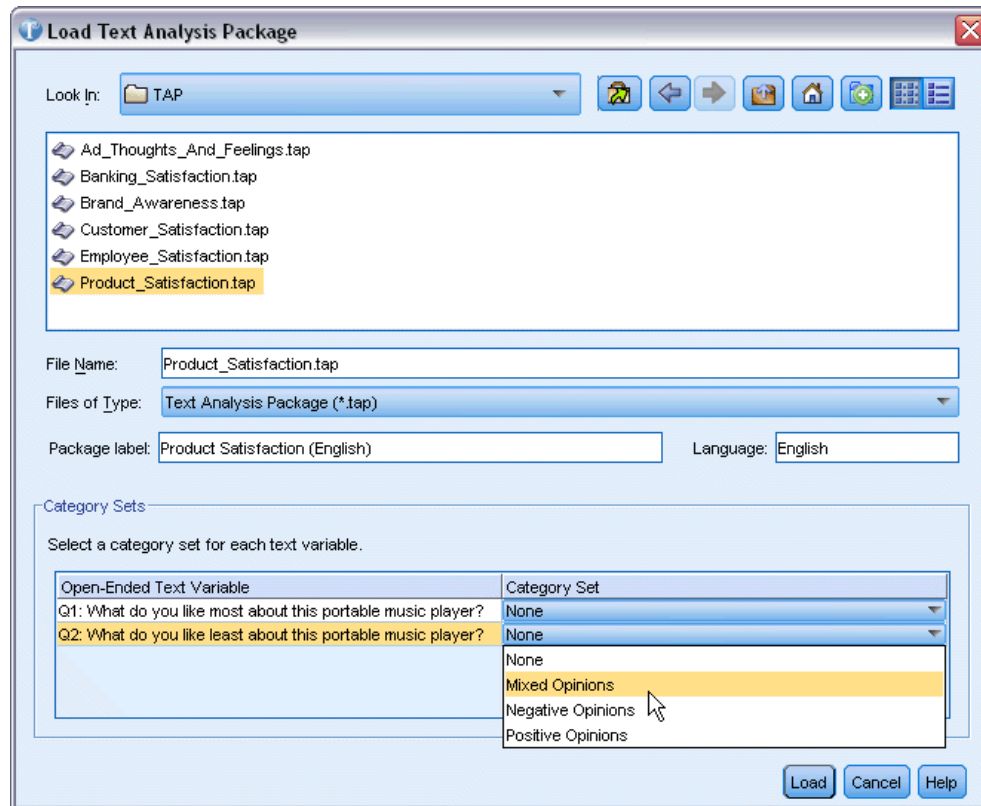
Note that if you have any templates in languages for which you have no license, a check box is displayed at the bottom of the dialog to enable you to hide the unlicensed language templates from display.

- ▶ Click Finish to start importing the data. If you choose this option, the resulting project will have the default linguistic resources and, after any extractions are performed, you can build your categories manually or using an automated technique. For more information, see the topic “Categorizing Text Data” in Chapter 6 on p. 91.

To select a text analysis package:

- ▶ To load a text analysis package, make sure the Text Analysis Package option is selected and click Load. The Load Text Analysis Package dialog appears.

Figure 3-9
Load text analysis package



- ▶ In the dialog, select the TAP you want to use. Only those packages stored in the default <installation_directory>\TAP directory appear directly in the list. The fields below update with the specific details for the selected TAP.
- ▶ In the Category Sets table, you can assign a category set to each of the text variables. In the Category Set column, click the drop-down list in each cell to choose an available category set. If you select None, then you will have no categories for that text variable until you create them later.
- ▶ Click OK. The dialog closes and the wizard now shows the new TAP you selected. After selecting the TAP and any category sets, the wizard is finished and in moments you can see your records coded into the prebuilt categories. From there, you can export the results or use the categories as a starting point for your analysis.
- ▶ Click Finish to close the dialog box and create the project. Once finished, the application automatically opens the Question view for the first open-ended text question in your project. If you chose to extract, an extraction progress dialog appears and it may take some time for the extraction process to complete. You can now begin to analyze your questions. To switch to a different question, from the menus choose View > Question.

Using Text Analysis Packages

A text analysis package, also called a TAP, serves as a template for text response categorization. Using a TAP is an easy way for you to categorize your text data with minimal intervention since it contains the code frame and the linguistic resources needed to code a vast number of records quickly and automatically. Using the linguistic resources, text data is analyzed and mined in order to extract key concepts. Based on key concepts and patterns found in the text, the records can be categorized into the category set you selected in the TAP. You can make your own TAP or update one.

A TAP is made up of the following elements:

- **Category Set(s).** A category set is essentially made up of a predefined categories, category codes, descriptors for each category, and lastly, a name for the whole category set. Descriptors are linguistic elements (concepts, types, patterns, and rules) such as the term *cheap* or the pattern *good price*. Descriptors are used to define a category so that when the text matches any category descriptor, the record is put into the category.
- **Linguistic Resources.** Linguistic resources are a set of libraries and advanced resources that are tuned to extract key concepts and patterns. These extraction concepts and patterns, in turn, are used as the descriptors that enable records to be placed into a category in the category set.

You can make and update text analysis packages.

After selecting the TAP and choosing a category set to each text variable in the New Project Wizard, IBM® SPSS® Text Analytics for Surveys can extract and categorize your records. From there, you can either export the results or continue fine-tuning the categorization until you get the results you want.

Note: TAPs can be created and used interchangeably between SPSS Text Analytics for Surveys and IBM® SPSS® Text Analytics.

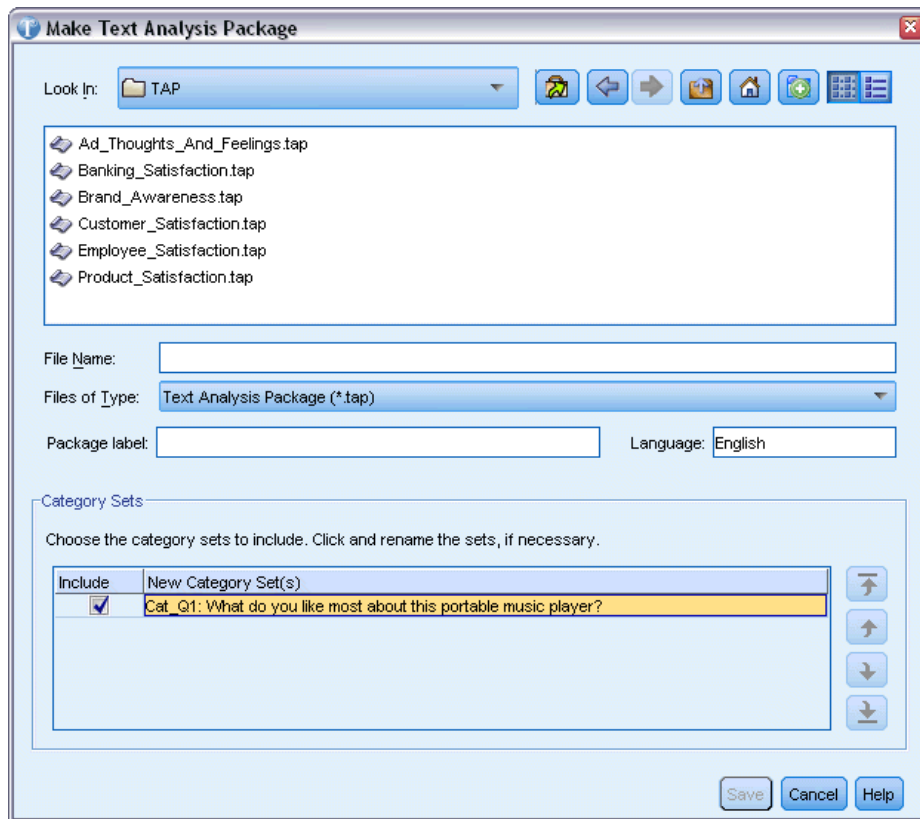
Making Text Analysis Packages

Whenever you have a project with at least one category and some resources, you can make a text analysis package (TAP) from the contents of the open project. The set of categories and descriptors (concepts, types, rules or TLA pattern outputs) in each question can be made into a TAP along with all of the linguistic resources open in the resource editor.

You can see the language for which the resources were created. The language is set in the Advanced Resources tab of the Resource Editor.

Important! If your categories contain text matches, forced records, or flags, those will not be saved into category sets since they are data-specific and almost always unusable on other data. However, labels and category codes are saved.

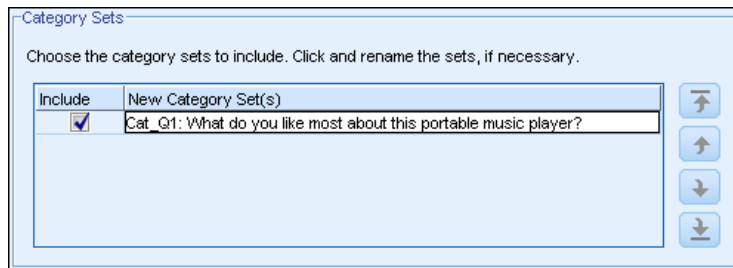
Figure 3-10
Make Text Analysis Package dialog



To Make a Text Analysis Package

- ▶ From the menus, choose File > Text Analysis Packages > Make Package. The Make Package dialog appears.
- ▶ Browse to the directory in which you will save the TAP. By default, TAPs are saved into the \TAP subdirectory of the product installation directory.
- ▶ Enter a name for the TAP in the File Name field.
- ▶ Enter a label in the Package Label field. When you enter a file name, this name automatically appears as the label but you can change this label. If you save a TAP in this default directory, the package label will appear as an option in the drop-down list in the New Project wizard.
- ▶ To exclude a category set from the TAP, unselect the Include checkbox. Doing so will ensure that it is not added to your package. By default, one category set per question is included in the TAP. There must always be at least one category set in the TAP.
- ▶ Rename any category sets. The New Category Set column contains generic names by default, which are generated by adding the Cat_ prefix to the text variable name. A single click in the cell makes the name editable. Enter or a click elsewhere applies the rename. If you rename a category set, the name changes in the TAP only and does not change the variable name in the open project.

Figure 3-11
Renaming category sets

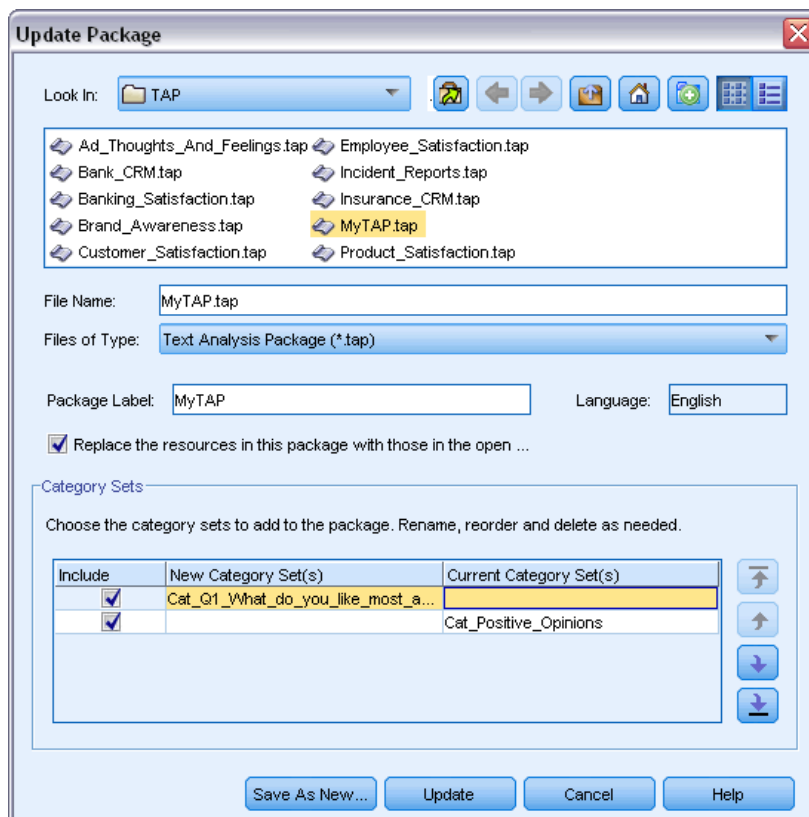


- ▶ Reorder the category sets if desired using the arrow keys to the right of the category set table.
- ▶ Click Save to make the text analysis package. The dialog box closes.

Updating Text Analysis Packages

If you make improvements to a category set, linguistic resources, or make a whole new category set, you can update a text analysis package (TAP) to make it easier to reuse these improvements later. To do so, you must be in the open project containing the information you want to put in the TAP. When you update, you can choose to append category sets, replace resources, change the package label, or rename/reorder category sets.

Figure 3-12
Update Text Analysis Package dialog



To Update a Text Analysis Package

- ▶ From the menus, choose File > Text Analysis Packages > Update Package. The Update Text Analysis Package dialog appears.
- ▶ Browse to the directory containing the text analysis package you want to update.
- ▶ Enter a name for the TAP in the File Name field.
- ▶ To replace the linguistic resources inside the TAP with those in the current project, select the Replace the resources in this package with those in the open session option. It generally make sense to update the linguistic resources since they were used to extract the key concepts and patterns used to create the category definitions. Having the most recent linguistic resources ensures that you get the best results in categorizing your records. If you do not select this option, the linguistic resources that were already in the package are kept unchanged.
- ▶ To update only the linguistic resources, make sure that you select the Replace the resources in this package with those in the open session option and select only the current category sets that were already in the TAP.
- ▶ To include the new category set(s) from the open project into the TAP, select the checkbox for each category set to be added. You can add one, multiple or none of the category sets.

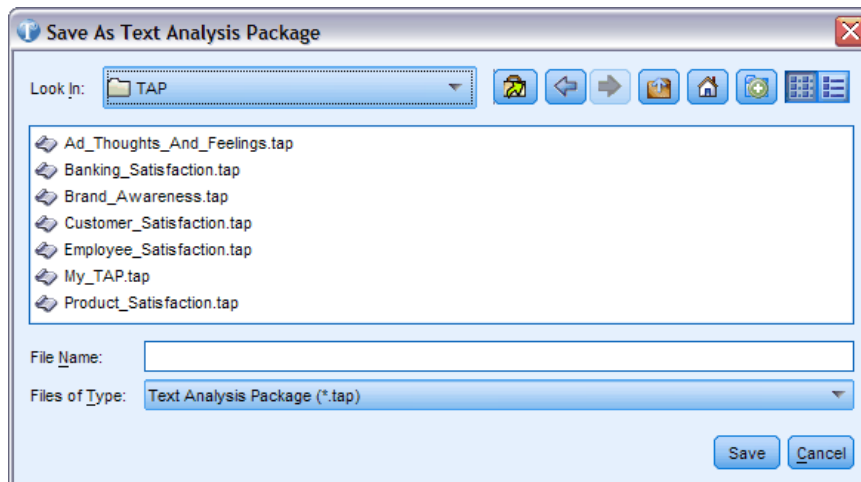
- ▶ To remove category sets from the TAP, unselect the corresponding Include checkbox. You might choose to remove a category set that was already in the TAP since you are adding an improved one. To do so, unselect the Include checkbox for the corresponding category set in the Current Category Set column. There must always be at least one category set in the TAP.
- ▶ Rename category sets if needed. A single click in the cell makes the name editable. Enter or a click elsewhere applies the rename. If you rename a category set, the name changes in the TAP only and does not change the variable name in the open project. If two category sets have the same name, the names will appear in red until you correct the duplicate.

Figure 3-13
Duplicate names

Include	New Category Set(s)	Current Category Set(s)
<input checked="" type="checkbox"/>	Cat_Q1: What do you like most about this p...	
<input checked="" type="checkbox"/>	Cat_Q2: What do you like least about this p...	
<input checked="" type="checkbox"/>		Cat_Q1: What do you like most about this p...
<input checked="" type="checkbox"/>		Cat_Q2: What do you like least about this p...

- ▶ To create a new package with the session contents merged with the contents of the selected TAP, click Save As New. The Save As Text Analysis Package dialog appears. See following instructions.
- ▶ Click Update to save the changes you made to the selected TAP.

Figure 3-14
Save As Text Analysis Package dialog



To Save a Text Analysis Package

- ▶ Browse to the directory in which you will save the TAP file. By default, TAP files are saved into the TAP subdirectory of the installation directory.
- ▶ Enter a name for the TAP file in the File name field.
- ▶ Enter a label in the Package label field. When you enter a file name, this name is automatically used as the label. However, you can rename this label. You must have a label. If you save a

TAP in this default directory, the package label will appear as an option in drop-down list in the New Project wizard.

- ▶ Click Save to create the new package.

Working with Projects

In IBM® SPSS® Text Analytics for Surveys, you will work with and categorize survey data. To do so, you will create projects in which you will build and store category definitions and the responses to which they correspond. A project can contain the following elements:

- **Survey data.** Text response variable for open-ended questions, a unique ID variable, and any optional reference variables. The survey data are not stored within the project; rather, they are read from the original data source when the project is opened.
- **Linguistic resources.** Proprietary and user-customized templates and libraries (synonyms, exclusions, and type dictionaries) used when extracting concepts and patterns from the text.
- **Extracted results.** Present after an extraction is performed, these are the key words and phrases identified and extracted from your response data. You will use these concepts to create your categories.
- **Categories.** Come from TAP category sets, manual creation and/or automated category building technique. Survey responses are assigned to these categories.

Opening Projects

You can return to an existing project by opening it. Only one project can be open at a time. If you attempt to open a project when one is already open, you will be prompted to save the other project first, if necessary.

When a project is opened, IBM® SPSS® Text Analytics for Surveys checks your linguistic resources to determine whether any public libraries are more recent than the ones in the project. If this is the case, you will be prompted about whether the libraries should be updated. You can then choose whether to keep your version and not update or to merge the updates into your project. For more information, see the topic “Updating Libraries” in Chapter 9 on p. 204.

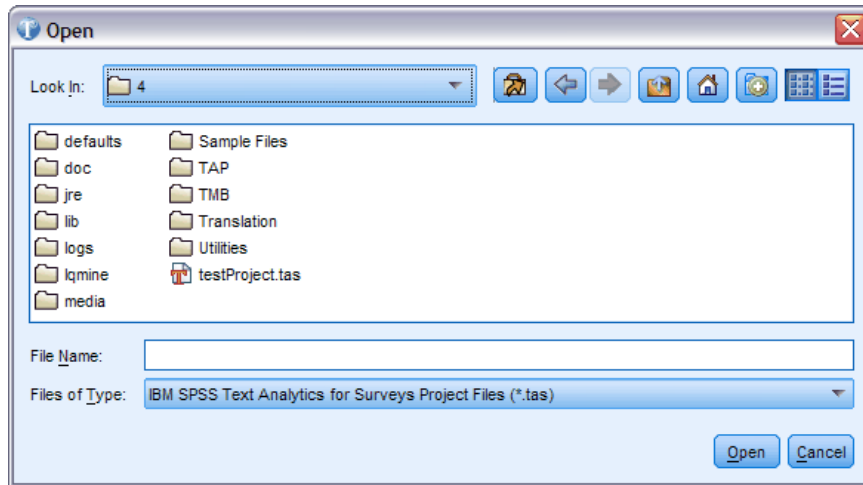
Important! SPSS Text Analytics for Surveys does not physically store the source data in its projects. Instead, a reference to that data on your machine is stored in the project. If someone changes any of the original imported variables in the data source, a warning appears that the data cannot be found. If this occurs, you must reimport the data and match the “new” questions with the original variable names to continue working with that project. In general, it is not recommended that you rename variables or column headers in your source data.

Note: The extraction results are saved within your projects unless you choose not to do so (Tools > Options). When you close a project, your category definitions are saved, but the Extraction Results pane will be cleared. When you open that project, you must run an extraction if you would like to continue categorizing your responses. The existing category definitions display a question mark (?) instead of a response count. After reextracting, the response counts will reappear.

To Open a Project

- ▶ From the menus choose File > Open Project. The Open dialog box opens.

Figure 4-1
Open dialog box



- ▶ From the list, select the directory and the name of the project that you want to open. You cannot sort the details in this dialog box, such as file size and date.
- ▶ Click OK to open the project in the main window.
- ▶ If your project contains data from a password-protected database, you are prompted for the password each time you open this project.
- ▶ If your project contains is from a previous version of the product, you will be prompted to convert your resources to the new format. This implies that after you save your project, you will not be able to open the project in an earlier version of this product.
- ▶ If any public libraries in your project have changed since you last opened the project, an alert will notify you of this change.

Important! Whenever you open a project, the corresponding data set is opened. If that data cannot be found, an error message appears. In order to continue working with your data, you must reimport the data. For more information, see the topic “Changing Data Sources” on p. 61.

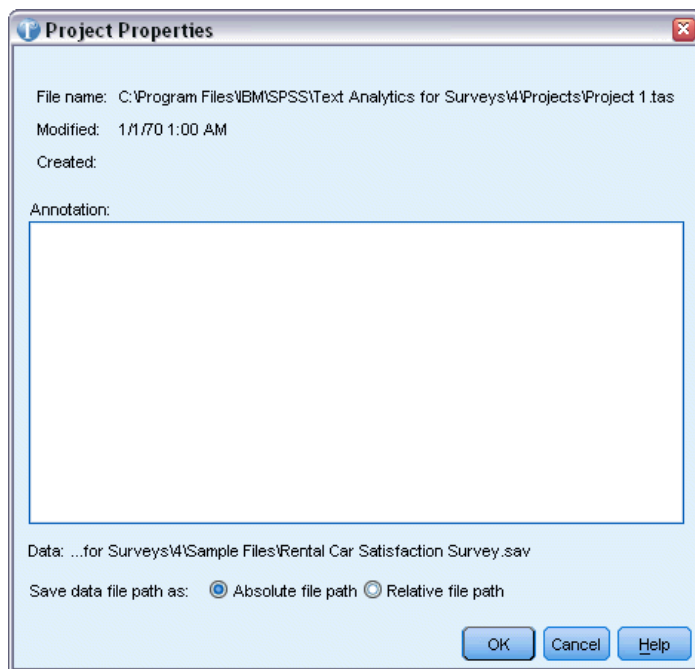
Editing Project Properties

By default, every new project is called *Project 1*. You can review the basic properties of your project as well as add or modify an annotation for your project.

To Edit Project Properties

- ▶ From the menus choose File > Project Properties. The Project Properties dialog box opens.

Figure 4-2
Project Properties dialog box



- ▶ If desired, enter a comment or description for the project in the Annotation text box.
Note: The name of the data file is shown in this dialog box. Since you are creating a project and have not yet imported data, the filename is not known. After importing, only the last 60 characters in the data filename will appear here. If you have a longer name, you can hover your mouse over the name to display the full name.
- ▶ Click OK to accept the new properties. The dialog box closes and the project properties are applied.

Viewing Project Data

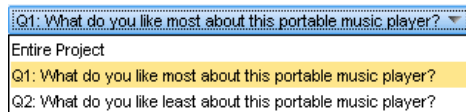
Once you have imported the data file, the Question view for the first open-ended text question in your project appears. However, you may want to look at all of the data that you imported. You can do so in the Entire Project view, which offers a comprehensive view of your data. To access this view, choose View > Entire Project from the menu. In this view, you can:

- Review the contents of all imported variables.
- Assign values and labels to the variables.
- Change the variable types.
- Sort the variables.
- Copy data from contiguous cells and paste them into other applications.
- Resize the variable columns.

Important! The data you imported into the project is read-only; you cannot edit this data from within IBM® SPSS® Text Analytics for Surveys.

You can then begin to extract concepts from these responses, with which you will create your categories. For more information, see the topic “Categorizing Text Data” in Chapter 6 on p. 91.

Figure 4-3
Accessing the Entire Project view



Sorting Variables

You can sort your data in the Entire Project view alphabetically or by length of data.

To Sort The Data in the Entire Project View

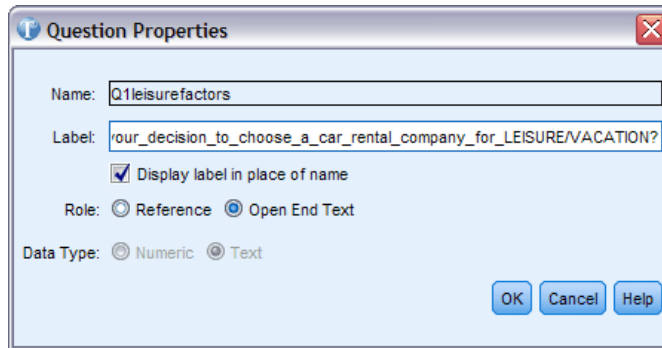
- ▶ Select the column that you want to sort and right-click the column title to open a context menu.
- ▶ Select the sort option that you want from the following choices:
 - **Natural Sort.** Results sort as they were read during import.
 - **Sort Ascending: A–Z.** Results sort alphabetically beginning with empty cells, numbers, and then A to Z.
 - **Sort Descending: A–Z.** Results sort alphabetically beginning with Z to A, numbers, and then empty cells.
 - **Sort Ascending: length.** Results sort by length, with the shortest responses appearing at the top.
 - **Sort Descending: length.** Results sort by length, with the longest responses appearing at the top.

Editing Variable Properties

While defining your data during the import process, you are asked to identify the variable representing the unique IDs, the variables representing the questions that you want to analyze, and, if applicable, any reference variables that you would like to include. After the data are imported, you may want to add information to the properties for these variables or change their role in the project. For example, you might want to analyze a variable that you imported as a reference variable. You can change the following variable properties:

- Add or change a variable name or label.
- Change a reference variable to an open-ended text variable.
- Change an open-ended text variable to a reference variable.
- Change the data type of an ID or reference variable.

Figure 4-4
Reference Properties dialog box



To Edit Variable Properties

- ▶ In the Entire Project view, select the column for the variable whose properties you want to modify and right-click the column title to open a context menu.
- ▶ Choose Properties from the menu. The Properties dialog box opens.
- ▶ If desired, add or modify the variable name or label.
- ▶ To use the variable labels instead of the variable name in the product, select the option Display label in place of name.
- ▶ If desired, change the variable's role in the analysis to either Reference or Open-Ended Text. You cannot change the role of the ID variable. If you have begun to work on an open-ended text variable (or question) and change its role to a reference variable, the categorization work that you have done will be lost.
- ▶ Change the data type of the variable to either Text or Numeric.

Saving Projects

Whenever you close a project or end the session, you are prompted to save any changes, if necessary. Projects are saved into files with the *.tas file extension.

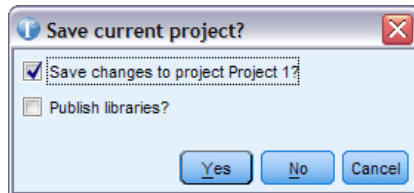
To Save Projects

- ▶ From the menus choose File > Save Project. The project is saved.

To Save A Project When Exiting

- ▶ When you close a project, a dialog box opens, asking whether you want to save the changes you made to the project and whether you want to (re)publish the libraries.

Figure 4-5
Save Current Project dialog box

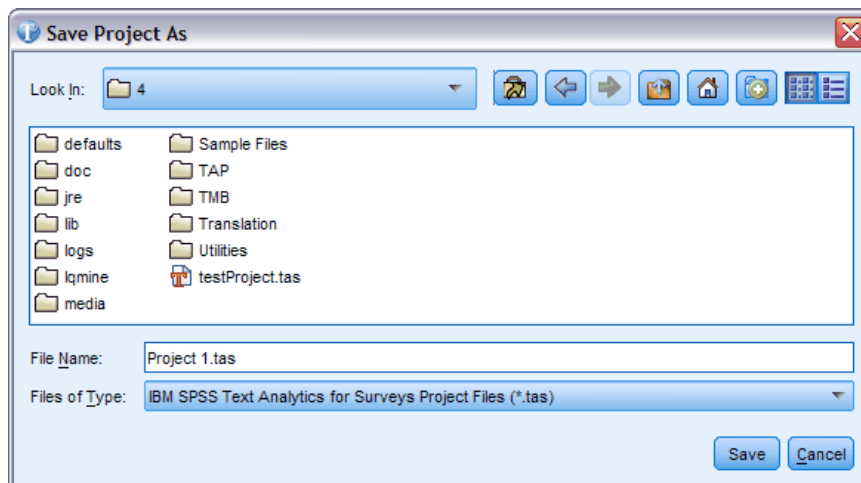


- ▶ Select Save changes to project.
- ▶ If you want to publish libraries for later reuse or to update public versions, select Publish libraries as well. If there are no libraries in need of publishing, this option is disabled. For more information, see the topic “Publishing Libraries” in Chapter 9 on p. 204.
- ▶ Click Yes to save. If you elected to publish libraries, another dialog box opens. For more information, see the topic “Sharing Libraries” in Chapter 9 on p. 202.

To Save As Another Project Name

If you receive an alert for a duplicate name or if you choose to save the project with a different name, the Save Project As dialog box opens.

Figure 4-6
Save Project As dialog box



- ▶ Enter the new, unique name for the project in the File Name text box.
- ▶ Click Save to save the new name.

Exporting Categorization Results

In some cases, the creation of categories may be the endpoint of your analysis. Simply knowing the major themes expressed by the respondents, and how many respondents mentioned each theme, may be adequate for the purposes of text analysis. However, often you may want to perform reporting or further analysis on the categories, such as creating tables and graphs to

display the results. You may even want to use other variables from the questionnaire to further characterize the respondents in each category or even use the categories to study other responses.

If you want to be able to continue working with your new categorization results, you can export your categories in a text format for import into a quantitative analytic application, such as the IBM® SPSS® Statistics Base system. The resulting file contains the IDs for the responses as well as the category names and labels, but it does not contain the values for any reference variables or the open-ended responses.

Note: You can also generate summary graphs, such as a Top 5 Categories bar chart. These graphs, which are exported into HTML, can then be used in presentations. For more information, see the topic “Exporting Summary Graphs” on p. 58.

Exported File Types

When you export, you can create one of several file types:

- SPSS Statistics files (*.sav). For more information, see the topic “Exporting to IBM SPSS Statistics or IBM SPSS Data Collection” on p. 54.
- Microsoft Excel files (*.xls / *.xlsx). For more information, see the topic “Exporting to Microsoft Excel” on p. 56.
- IBM® SPSS® Data Collection. For more information, see the topic “Exporting to IBM SPSS Statistics or IBM SPSS Data Collection” on p. 54. Also refer to the Data Collection Development Library under *Data Collection Data Model*.

Dichotomies versus Categories Output

Text data that have been coded with IBM® SPSS® Text Analytics for Surveys form a multiple-response set, since each respondent can give more than one response and can be assigned to more than one category for a single question. This means that the data must be coded in a special format when exported. Two different output formats are available when exporting: **dichotomies** and **categories**.

Dichotomies. The results center on category membership flags for each response ID. For each category in the data, each respondent (by ID) either belongs or does not belong to the category using a binary flag, which is coded either true or false. The data are structured in a table format, with the ID in the left column and one column for each category. This data type allows an unlimited number of categories per response. If there are 10 categories, there will be 10 new variables.

Categories. The results center on the set of categories to which a response belongs. For each response in the data, each category to which it is assigned appears as either a value (for SPSS Statistics) or the category itself (for Microsoft Excel). The category export data are structured in table format, with the ID in the left column, followed by one column per category to which at least one response belongs. These columns do not represent a particular category but rather a slot to record an assigned category code. For each response, each category code to which the response belongs is stored in a separate slot. The response with the maximum number of categories assigned to it determines the number of variables to be created. If there are 10 categories but no respondent is coded with more than 4 categories, then 4 variables will be needed to represent the categories.

- **For SPSS Statistics/Data Collection.** For each response ID in the data, each category to which it is assigned appears as a separate value from 1 to N , where N is the highest category code value. If you did not assign codes in the Code Frame Manager, then the codes were assigned automatically when the category was created. If a respondent is assigned to less than the maximum number of categories, the remaining unused variables will be coded with the SPSS Statistics system-missing value (a period).
- **For Microsoft Excel.** For each response ID in the data, each category to which it is assigned appears as either the category name or category label, depending on what you are using in the product interface. If a respondent is assigned to less than the maximum number of categories, the remaining unused variables will be coded with a blank in Microsoft Excel.

Exporting to IBM SPSS Statistics or IBM SPSS Data Collection

Once your responses are categorized, you will probably want to analyze your results using statistical procedures. IBM® SPSS® Text Analytics for Surveys allows you to create a data file that is formatted for use within different products—these instructions are for exporting for use within IBM® SPSS® Statistics (statistical analysis program) and various IBM® SPSS® Data Collection products. SPSS Text Analytics for Surveys will automatically create the multiple-response variable in your exported file. The exact format of the file depends on the data type you select—dichotomies or categories.

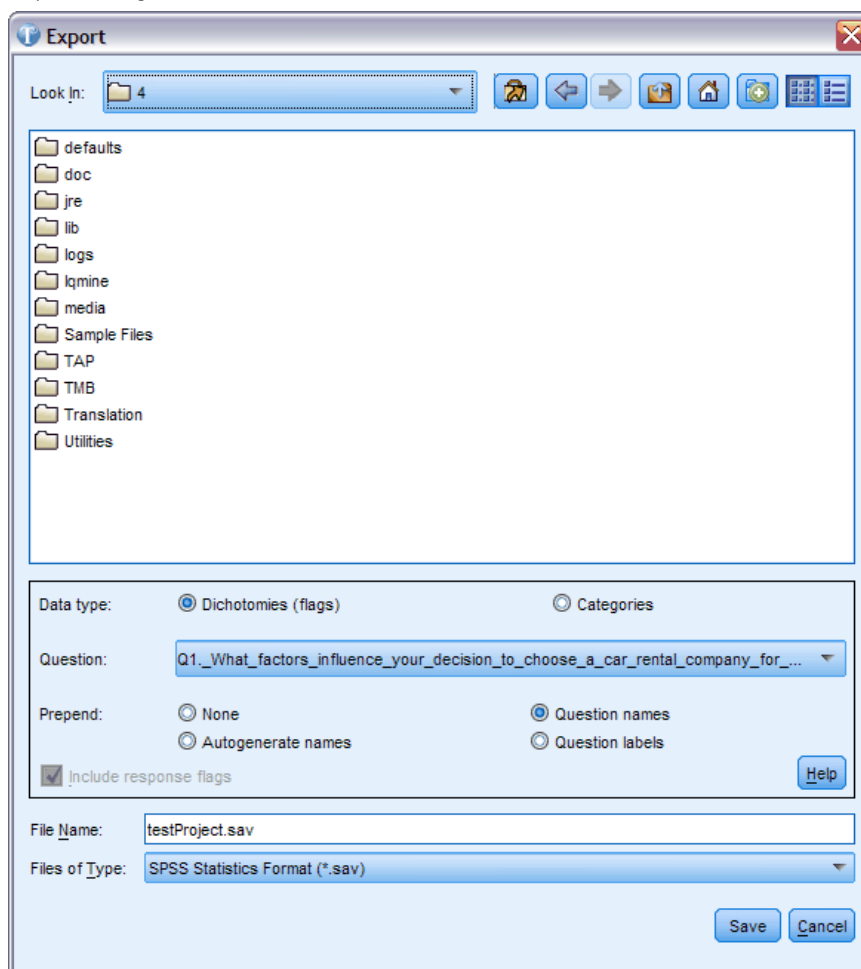
Note: The resulting file contains the IDs for the responses as well as the category names and labels, but it does not contain the values for any reference variables or the open-ended responses.

SPSS Statistics only. For the output, if your data set contains missing data or cases in which a respondent did not answer a particular question, the application assigns the system-missing value to these cases. SPSS Statistics files exported by SPSS Text Analytics for Surveys are not supported by SPSS Statistics versions prior to 7.5.

To Export Data

- ▶ From the File > Export Results menu choose one of the following options to open the Export dialog box:
 - SPSS Statistics File
 - Data Collection File

Figure 4-7
Export dialog box for IBM SPSS Statistics file formats



- ▶ From the Save In drop-down list, select the drive and folder in which you want to save the file.
- ▶ Select a Data Type option. For more information, see the topic “Exporting Categorization Results” on p. 52.
 - Dichotomies.
 - Categories. This option is not available for the Data Collection data file, and Dichotomies is selected by default.
- ▶ From the Question drop-down list, select the question that you want to export. You can choose whether to export the categorization results for a single question or for the entire project. If you want to export each question separately, you must select and export each question one at a time. Or, select Entire Project to export the results for all open-ended questions.
- ▶ Select a Prepend option to designate a prefix when exporting category names for the *entire project*. This option is most useful when exporting the data for multiple questions. Prepending adds a prefix to the original variable label or category name and ensures that you have no duplicates

when combining the results for multiple questions when exporting the entire project. Choose from the following:

- None. As the option implies, no prefix is added.
 - Question names. Prefixes outputted category variable names (either the category name or the category label depending on what you were using in your project) with the open-ended text variable (question) name. The question variable name comes from the original data source. If the outputted category variable name doesn't meet variable-naming conventions or exceeds 40 characters, then default names are created (per the Autogenerate option).
 - Autogenerate names. Automatically prefixes category names with *Q1*, *Q2*, *Q3*, and so on. *Q1* refers to the first question you export, and so on.
 - Question labels. Prefixes outputted category variable names (either the category name or the category label depending on what you were using in your project) with the open-ended text variable (question) label. The question variable name comes from the original data source. If the outputted category variable name doesn't meet variable-naming conventions or exceeds 40 characters, then default names are created (per the Autogenerate option).
- ▶ If you have response flags in your data, you can choose whether to export them as well. To export response flags, select that option. For more information, see the topic “Flagging Responses” on p. 73.
 - ▶ In the File Name text box, select the default project name that appears, or enter another name for this file.
 - ▶ Click Save to export the results.

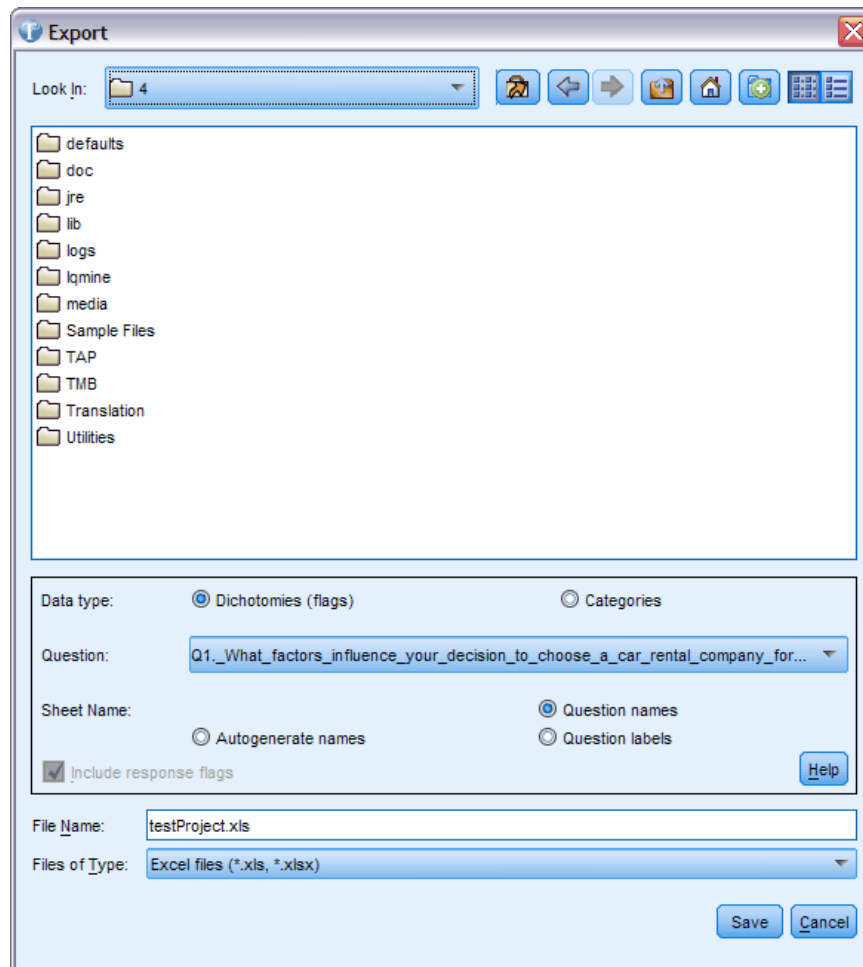
Exporting to Microsoft Excel

Once your responses are categorized, you will probably want to analyze your results using statistical procedures. IBM® SPSS® Text Analytics for Surveys allows you to create a data file that is formatted for use within different products. The following instructions are for exporting into an Microsoft Excel format. SPSS Text Analytics for Surveys will automatically create the multiple-response variable in your exported file. The exact format of the file depends on the data type you select—dichotomies or categories. The resulting file contains the IDs for the responses as well as the category names and labels, but it does not contain the values for any reference variables or the open-ended responses.

To Export Data

- ▶ From the menus, choose File > Export Results > Microsoft Excel File. The Export dialog box opens.

Figure 4-8
Export dialog box for Microsoft Excel files



- ▶ From the Save In drop-down list, select the drive and folder in which you want to save the file.
- ▶ Select a Data Type option. For more information, see the topic “Exporting Categorization Results” on p. 52.
 - Dichotomies.
 - Categories. This option is not available for the IBM® SPSS® Data Collection data file, and Dichotomies is selected by default.
- ▶ From the Question drop-down list, select the question that you want to export. You can choose whether to export the categorization results for a single question or for the entire project. If you want to export each question separately, you must select and export each question one at a time. Or, select Entire Project to export the results for all open-ended questions.

- ▶ Select a worksheet naming option to designate how each worksheet generated in the exported Microsoft Excel file should be named. Choose from the following:
 - Question names. Uses the open-ended text variable (question) name as the worksheet name. The question name comes from the original data source. If the outputted category variable name doesn't meet variable-naming conventions or exceeds 40 characters, then default names are created (per the Autogenerate option).
 - Autogenerate names. Automatically names the worksheets with *Q1*, *Q2*, *Q3*, and so on. *Q1* refers to the first question you export, and so on.
 - Question labels. Uses the open-ended text variable (question) label, if one exists, as the worksheet name. If the outputted category variable name doesn't meet variable-naming conventions or exceeds 40 characters, then default names are created (per the Autogenerate option).

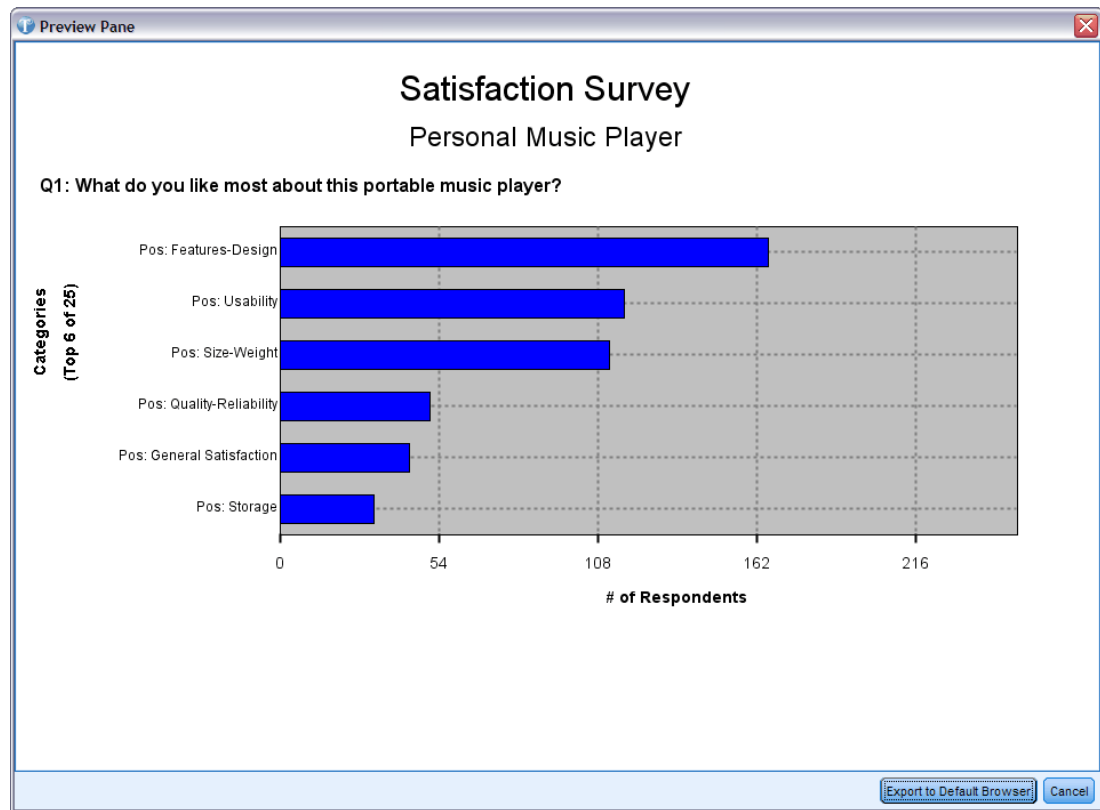
- ▶ If you have response flags in your data, you can also choose whether to export them. To export response flags, select that option. For more information, see the topic "Flagging Responses" on p. 73.

- ▶ In the File Name text box, select the default project name that appears, or enter another name for this file.

Exporting Summary Graphs

When you are done working with your categories and data, you can export graphical summary reports in order to share your analysis results and findings with others. The output produces one bar chart per question. You can choose the number of top categories to use in each graph so that you can visually present the top 5 or top 10 categories for a given question. The graph can be exported to your default browser from which you can save the image for use in other products or presentations.

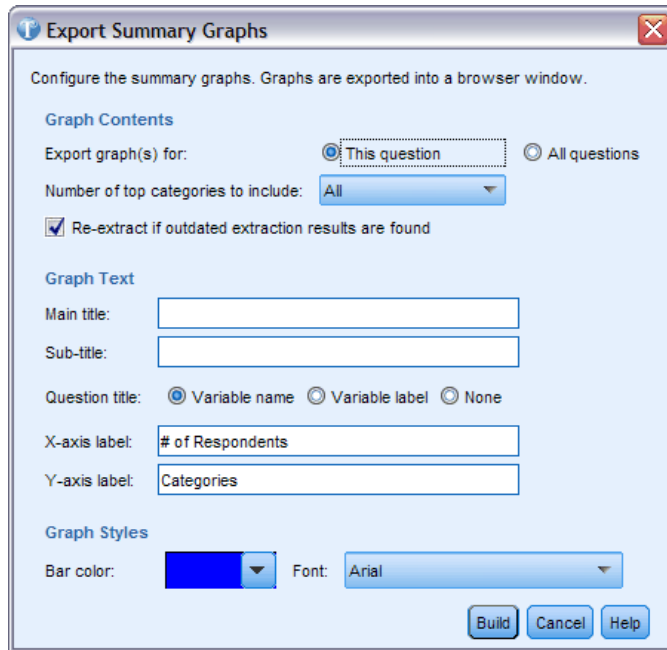
Figure 4-9
Sample summary graphs in a browser window



To Export Summary Graphs

- ▶ From the menus, choose Categories > Export Summary Graphs. The Export Summary Graphs dialog box opens.
- ▶ Configure your summary graph using the options described in this topic.
- ▶ Click Build to generate the graph and display it in a Preview pane.
- ▶ Click Export to Default Browser to see the graph in a browser window.

Figure 4-10
Export Summary Graphs dialog box



Report Options

Export graph(s) for. Choose whether you want to generate a summary graph for all of the questions in your project or only for the currently selected question.

Number of top categories to include. Select the maximum number of categories to display in the graph. Those categories with the greatest number of records are used first.

Re-extract if outdated extraction results are found. Select this option to force a re-extraction before generating the graph if the extraction results are not up to date.

Main title. Enter a main title for your graphs. For example, this could be the name of your survey.

Sub-title. Enter a sub-title for your graphs. For example, this could be the name of the company or the year of the survey.

Question title. To help you identify each chart, the title is derived from the question. Choose whether to use the question's variable name, label, or no name at all.

X-axis label. Define a label for the X-axis of the graphs. A label is proposed by default.

Y-axis label. Define a label for the Y-axis of the graphs. A label is proposed by default.

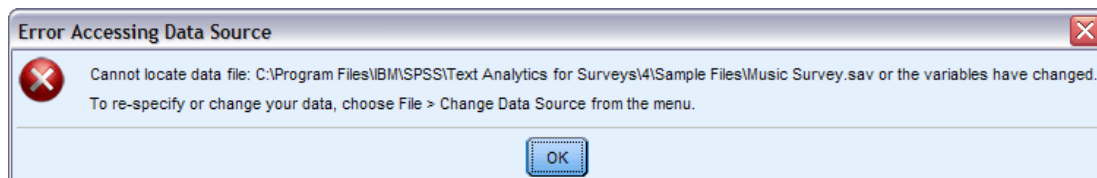
Bar color. Choose a color for the bars in the summary graph. This color applies to all of questions.

Font. Choose a font for the titles and labels in the graph.

Changing Data Sources

Whenever you open a project, the corresponding data set is opened. If that data cannot be found, an error message appears. Sometimes data cannot be found because they were moved to a different location, accidentally deleted by someone, or renamed. Alternately, you might want to switch data sources.

Figure 4-11
Error message for missing data

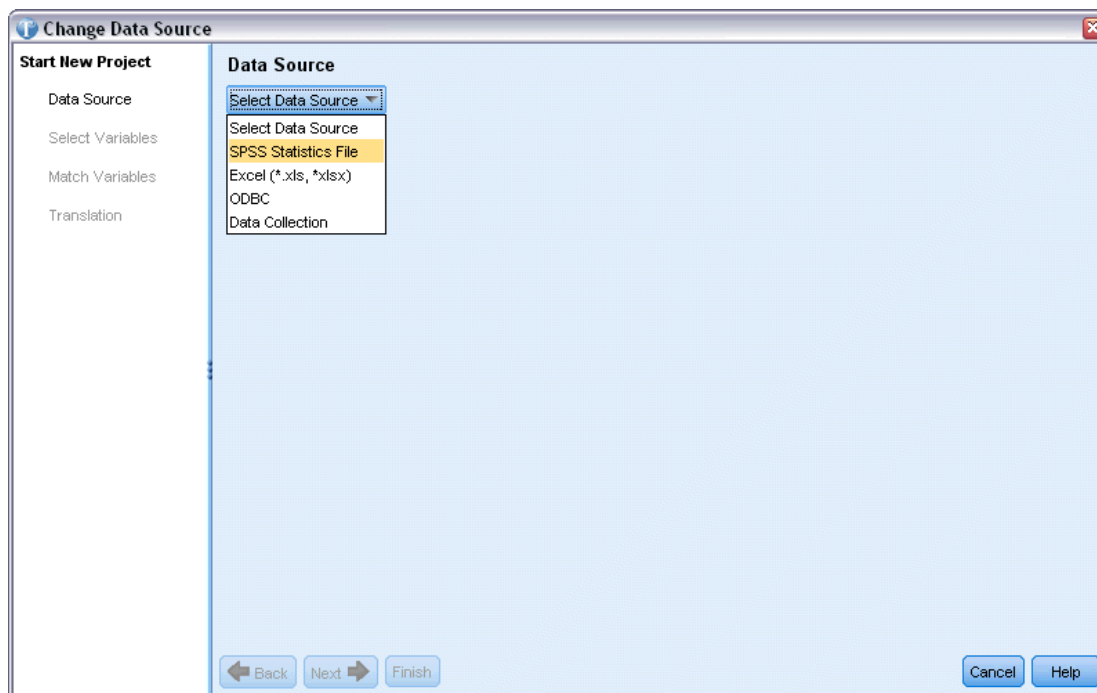


To continue working with your data, you must change the location to the proper data source. If any variable changes are found in the data, such as new variables, renamed variables, or missing variables, you will be asked to match the previously imported variables to the new ones.

To Change Your Data Source:

- ▶ When you receive this error message, click OK.
- ▶ From the menus choose File > Change Data Source. The Change Data Source wizard dialog appears.

Figure 4-12
Change Data Source wizard



Selecting Data Sources

When the wizard opens, you begin by selecting a data source. IBM® SPSS® Text Analytics for Surveys was optimized to process data sets of up to 10,000 records, although performance will vary based on the volume of text contained in these records. See the installation instructions for performance statistics and recommendations.

Important! An ID variable with a unique value for each record must be present in order to import the data.

You can choose one of the following data sources:

- **SPSS Statistics files** (*.sav).
- **Microsoft Excel files** (*.xls / *.xlsx).
- **ODBC** (Microsoft Open Database Connectivity protocol) database.
- **Data Collection** data model. This option is available only if you have the data model installed.

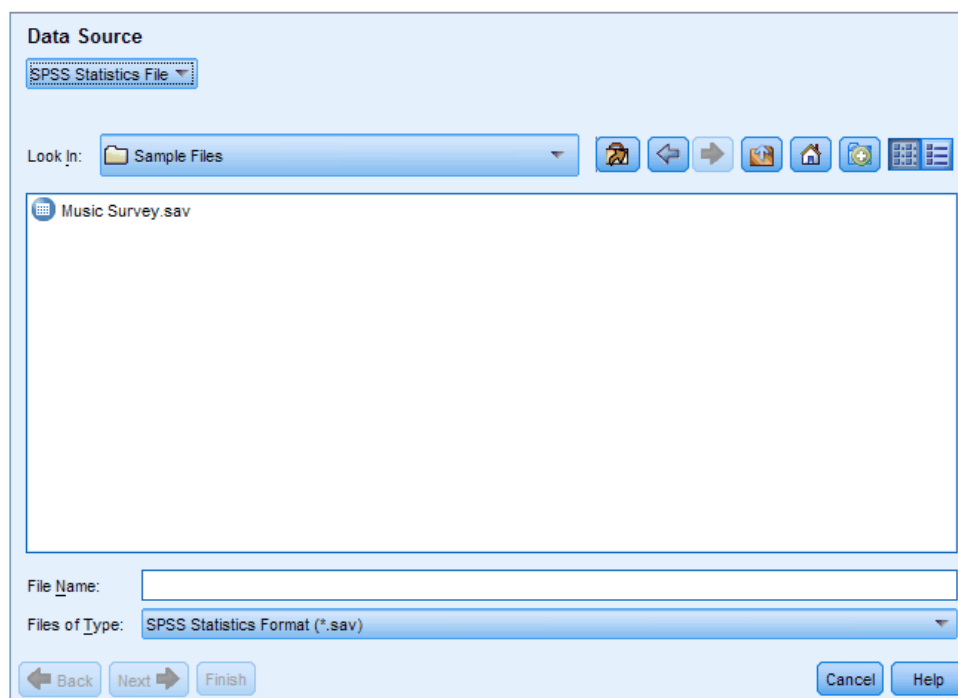
Using IBM SPSS Statistics Files

You can import an IBM® SPSS® Statistics (.sav) file into IBM® SPSS® Text Analytics for Surveys. An ID variable with a unique value for each record must be present in order to import the data.

Important! You cannot import SPSS Statistics (.sav) file with records exceeding 4000 characters.

Note: SPSS Text Analytics for Surveys was optimized to process data sets of up to 10,000 records, although performance will vary based on the volume of text contained in these records. See the installation instructions for performance statistics and recommendations.

Figure 4-13
Data source options for IBM SPSS Statistics files



To Get Data from IBM SPSS Statistics

- ▶ In the first screen of the wizard, select SPSS Statistics file from the drop-down list. The wizard displays the options for SPSS Statistics files.
- ▶ From the Look In drop-down list, select the drive and folder in which the file is located.
- ▶ Select the file from the list. It will appear in the File Name text box.
- ▶ Click Next to select variables. For more information, see the topic “Selecting Variables” on p. 32.

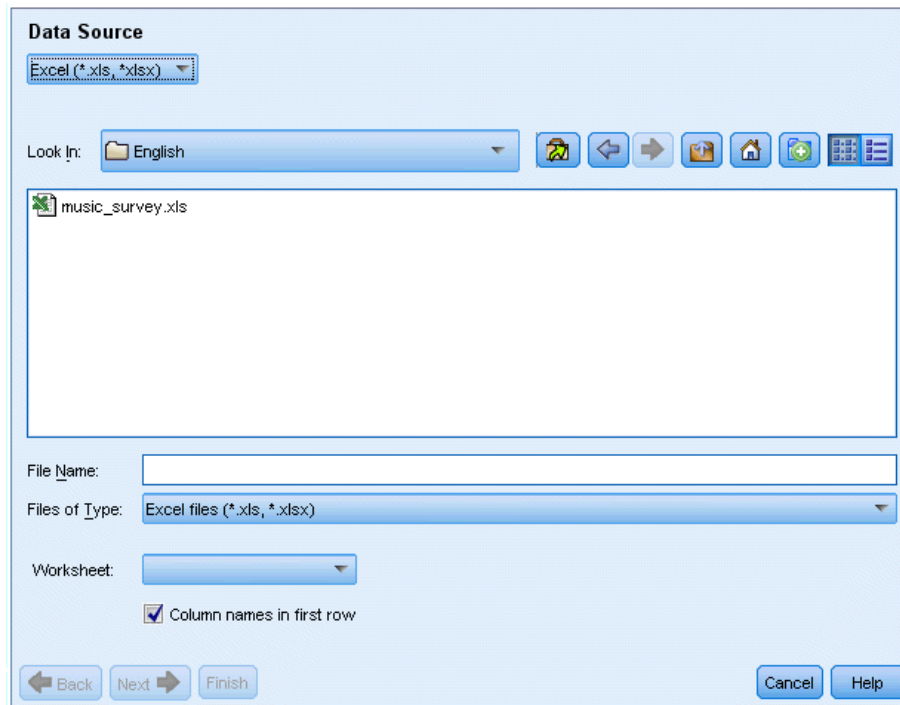
Using Microsoft Excel Files

You can import a Microsoft Excel (.xls / .xlsx) file into IBM® SPSS® Text Analytics for Surveys. An ID variable with a unique value for each record must be present in order to import the data.

Important! During the Microsoft Excel file import, you can select an option for Column Names in First Row. To use this option, the very first row of the worksheet must contain column names—not the row just above where the data begin. For example, if your data and column names begin on row 7, you must delete rows 1–6 before importing in order to import the file correctly.

Note: SPSS Text Analytics for Surveys was optimized to process data sets of up to 10,000 records, although performance will vary based on the volume of text contained in these records. See the installation instructions for performance statistics and recommendations.

Figure 4-14
Data source options for Microsoft Excel files



To Get Data from Microsoft Excel

- ▶ In the first screen of the wizard, select Excel from the drop-down list. The wizard displays the options for Microsoft Excel files.
- ▶ From the Look In drop-down list, select the drive and folder in which the file is located.
- ▶ Select the file from the list. It will appear in the File Name text box.
- ▶ Select the worksheet from the drop-down list. You can only import data from a single worksheet. To work with data on multiple worksheets, you must create multiple projects.
- ▶ If the first row of this worksheet contains the column headers, select Column Name in First Row. To use this option, the very first row of the worksheet must contain column names—not the row just above where the data begin. For example, if your data and column names begin on row 7, you must delete rows 1–6 before importing in order to import the file correctly. The application can use these (or a converted version if the column headings do not conform to IBM® SPSS® Statistics variable-naming conventions) as the variable names. If not, the application will use the spreadsheet column letters as identifiers.
- ▶ Click Next to select variables. For more information, see the topic “Selecting Variables” on p. 32.

Using Data through ODBC

Data from database sources, commonly databases, are easily imported into IBM® SPSS® Text Analytics for Surveys. Any database that uses Open Database Connectivity (ODBC) drivers can be read directly by the product after the proper drivers are installed on the machine on which SPSS Text Analytics for Surveys is installed. An ID variable with a unique value for each record must be present in order to import the data.

Note: SPSS Text Analytics for Surveys was optimized to process data sets of up to 10,000 records, although performance will vary based on the volume of text contained in these records. See the installation instructions for performance statistics and recommendations.

Figure 4-15
Data source options for ODBC

Data Source

ODBC

Source(DSN)

Name	Description
Visio Database Samples	Microsoft Access Driver (*.MDB)
MS Access Database	Microsoft Access Driver (*.mdb)
Excel Files	Microsoft Excel Driver (*.xls)
dBASE Files	Microsoft dBase Driver (*.dbf)

ODBC

User:

Password:

Table

SQL

Back Next Finish Cancel Help

To Use Via ODBC

- ▶ In the first screen of the wizard, select ODBC from the drop-down list. The wizard displays the options for ODBC.
- ▶ Specify the data source by selecting it from the list of registered ODBC sources or by typing the name into the Source (DSN) text box. If you need to register new data sources that do not appear in the list, click ODBC. This will open the ODBC Data Source Administrator, which is found on most Microsoft Windows computers. If it is not found, you cannot use the ODBC import. Consult the Microsoft Windows Help system for more information.
- ▶ If the data source is password protected, you must enter a user name and password. You will be required to do so each time you open the project, since, for security reasons, the user name and password are not stored in the project.

- ▶ Select your data in one of two ways: Table or SQL. You can select a table directly or use SQL commands to select data.
- ▶ Click Next to select variables. For more information, see the topic “Selecting Variables” on p. 32.

Using IBM SPSS Data Collection Data

To Import Via IBM SPSS Data Collection

- ▶ In the first screen of the wizard, select Data Collection from the drop-down list. The IBM® SPSS® Data Collection data model option is available only if you have the data model installed with another product.

Selecting Variables

After selecting the data source, the next step is to specify the variables to be imported. Three types of variables can be imported into a project.

Unique ID Variable (Required)

The ID variable is a unique numeric or string key that identifies each respondent. The data file does not need to be ordered by the unique ID variable to successfully read it. After being read into the program, the records can be sorted by various criteria. For more information, see the topic “Sorting Variables” on p. 50. This ID variable is required to import data. Each imported record (or case) must have a unique ID value.

Two situations will cause the import to fail:

- Duplicate ID values detected
- Records with blank ID values

Note: If a duplicate ID is detected and you have IBM® SPSS® Statistics installed on your computer, you can use the Identify Duplicate Cases procedure in that product to identify duplicates and then use the options to indicate which records should be retained (primary cases).

Open-Ended Text Variable(s) (Required)

The open-ended text variables represent the text responses to the question(s) in the survey. At least one of these variables is required to import data. These variables can be string or long-string variables in SPSS Statistics, columns containing general or text cells in Microsoft Excel, or text or note fields from databases. Each open-ended text variable will be analyzed separately. There is a 4,000-character limit on the size (width) of each text variable imported from a .SAV file.

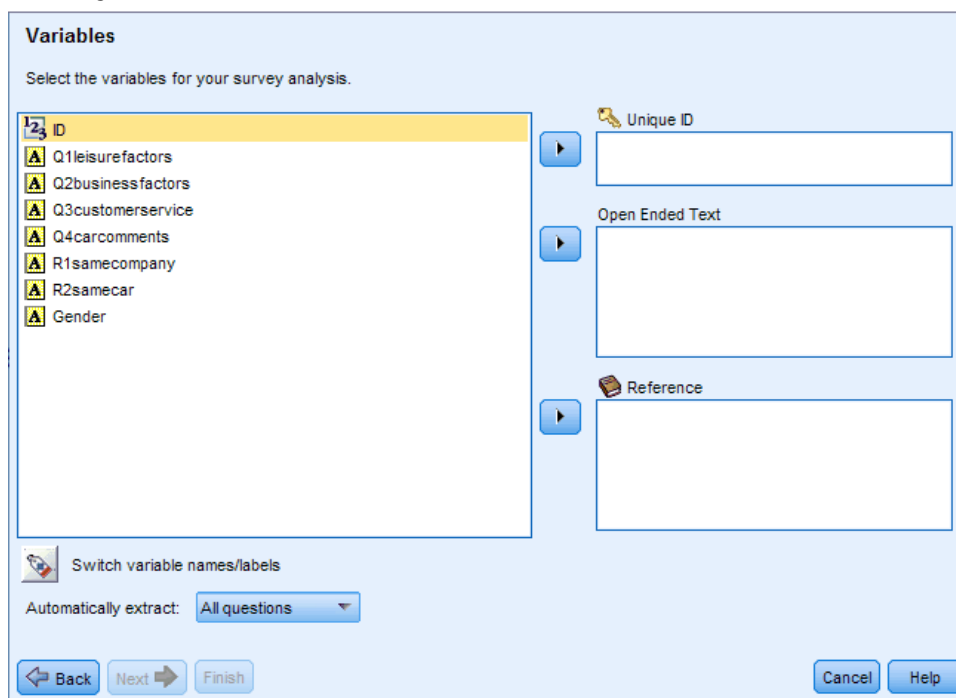
Reference Variable(s) (Optional)

The reference variables are additional, optional variables, generally categorical, that can be imported for reference purposes. Reference variables are not used in text analysis but provide supplemental information describing the respondent, which may aid understanding and

interpretation. Demographic variables are often included as reference variables, since they can contribute to understanding which terms or categories are being used by which groups of individuals. Examples are sex, department, occupation, and course of study (for student and training evaluations). You can view all of the reference variables after importing in the Entire Project view. You can also display reference variables in the Data pane of the Question view. Additionally, you can select reference variables in the bar chart in the visualization pane to be able to drill down to a subset of respondents.

Note: Reference variables read from an SPSS Statistics data file will have variable labels (if supplied) appearing as column headings and their value labels (if supplied) displaying in the cells of the Data pane.

Figure 4-16
Selecting variables



To Select Variables and Extraction Options

- ▶ From the list of available variables, select the variable that corresponds to the ID variable in your data set and click the arrow button to move it into the Unique ID box. The ID must be a unique number or alphanumeric string that distinguishes one record from another. If your data set contains duplicate IDs, an error message appears. In this case, you must clean your data before trying again.
- ▶ From the list of available variables, select one or more variables that correspond to the open-ended response variables and click the arrow button to move the variable(s) into the Open-Ended Text list. The variable(s) will each be imported as a separate question whose responses you will analyze and categorize.
- ▶ From the list of available variables, select one or more variables that correspond to the reference variables and click the arrow button to move the variable(s) into the Reference list. Reference

variables are not used by the automated category building techniques. However, you can view their content and use them to help you make informed decisions when categorizing your responses.

- ▶ To view the variable labels instead of the variable names, click the button below the variable list on the left.
- ▶ To change the extraction setting, make a selection in the drop-down list. By default, First question only is selected, which means that if you have selected more than one open ended text variable, the extraction process will start automatically for the first question after the wizard ends. Extraction can take some time with larger data sets. Therefore, you may choose to extract None or All questions depending on the time available.
- ▶ Click Next > once you have selected all of your variables.

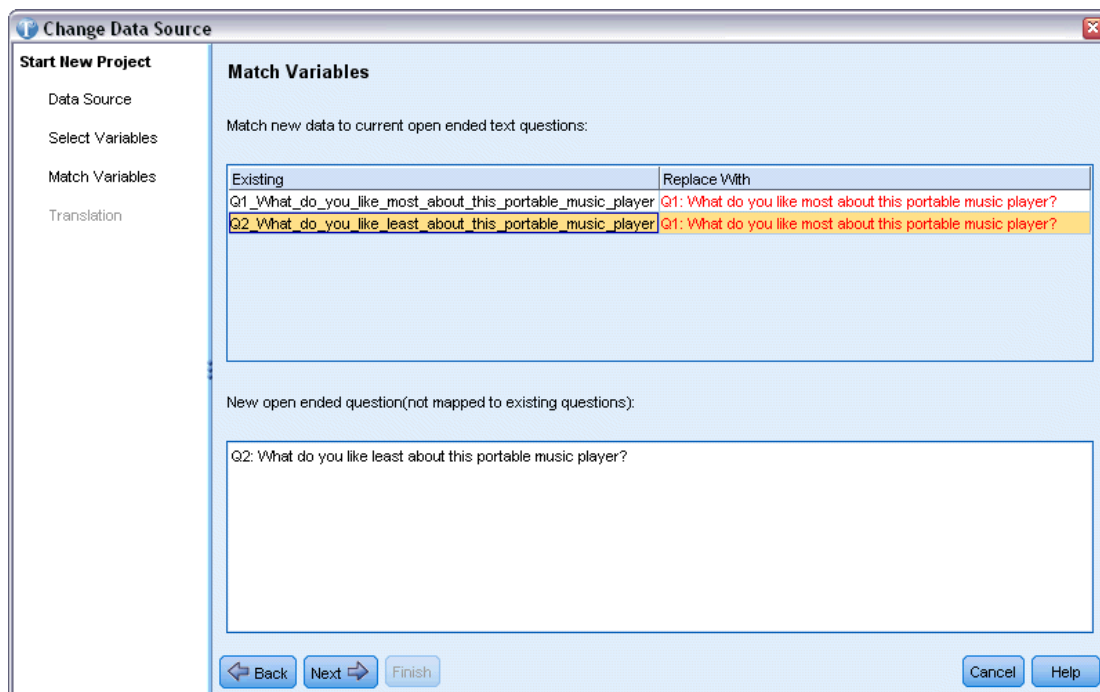
Matching Variables

After selecting variables in the previous step, IBM® SPSS® Text Analytics for Surveys attempts to match the previously imported variables to those you just selected. Matches are automatically proposed but you can match up the variables differently by clicking inside the Replace With column and choosing another variable.

If the new data file does not contain an open-ended text variable that existed in the project before, you can select NONE from the list, and any data associated with the old question will be discarded from the project.

Any variables in the new data file that are not mapped to the existing project appear in the New Open-Ended Question list at the bottom of the dialog box. After you change data sets, these remaining variables will appear as new questions in your project.

Figure 4-17
Match Contents to Existing Project dialog box



To match variables to existing ones:

- ▶ While the product attempts to match your new variables to the ones that were previously found in the data file, you can change how the variables are matched by clicking in the Replace With column and selecting the variable match. If you selected more variables in the previous step than specified in the Replace With column, then they will appear in the New open ended questions (not mapped to existing questions). After you change data sets, these remaining variables will appear as new questions in your project.
- ▶ If the existing variable has no match in the new data set, choose NONE and the data from this existing question will be discarded for you.
- ▶ If your project contains any responses that have been forced into or out of categories or any flags, you will be prompted to keep or discard these response ID-specific results. Typically, if you are importing different data (new questions or different respondents, for example), you will generally want to discard this information so as not to produce false results. If you are importing the same data file, you will generally want to keep this information since the response IDs would match up to the old data.

Translating into English

If you are working with non-English source text, you can connect to Language Weaver to translate into English. Translation is only available into English. You must have Language Weaver properly configured and with connections defined to translate. For more information, see the topic “Options: Translation Tab” in Chapter 2 on p. 21.

Figure 4-18
Translation options



To Translate Into English

- ▶ To translate the text data from a licensed language into English, select the Translate into English checkbox.
- ▶ From the Language Pair Connection list, select the connection for the Language Weaver language pair you want to use. If you have Language Weaver configured on your local machine, those language pairs will automatically appear in this list. You can add, change, or test the online services connection in the Translation tab of the Options dialog. For more information, see the topic “Options: Translation Tab” in Chapter 2 on p. 21.
- ▶ Specify the desired Translation accuracy. Choose a value of 1 to 3 indicating the level of speed versus accuracy you want. A lower value produces faster translation results but with diminished accuracy. A higher value produces results with greater accuracy but increased processing time. To optimize time, we recommend beginning with a lower level and increasing it only if you feel you need more accuracy after reviewing the results.
- ▶ If you have previously created custom dictionaries, held by Language Weaver, you can use them in connection with the translation. To choose a custom dictionary, select the Use custom dictionary checkbox and enter the Dictionary name. To use more than one dictionary, separate the names with a comma.
- ▶ In the New Project Wizard, click Next > to begin selecting categories and resources. For more information, see the topic “Selecting Categories and Resources” in Chapter 3 on p. 36.

- ▶ In the Change Data Set Wizard, click Finish to complete the data set change and to start the translation process.

To skip translation:

- ▶ Unselect the Translate into English option.
- ▶ In the New Project Wizard, click Next > to begin selecting categories and resources. For more information, see the topic “Selecting Categories and Resources” in Chapter 3 on p. 36.
- ▶ In the Change Data Set Wizard, click Finish to complete the data set change.

Updating Data

As you work with your project data, you might change the original data source. For example, you might add or remove records. You can update and refresh the data using the Update Data feature. However, if you have changed the variable names or the filename, for example, you will have to reimport your data completely. For more information, see the topic “Changing Data Sources” on p. 61.

To update and refresh your data:

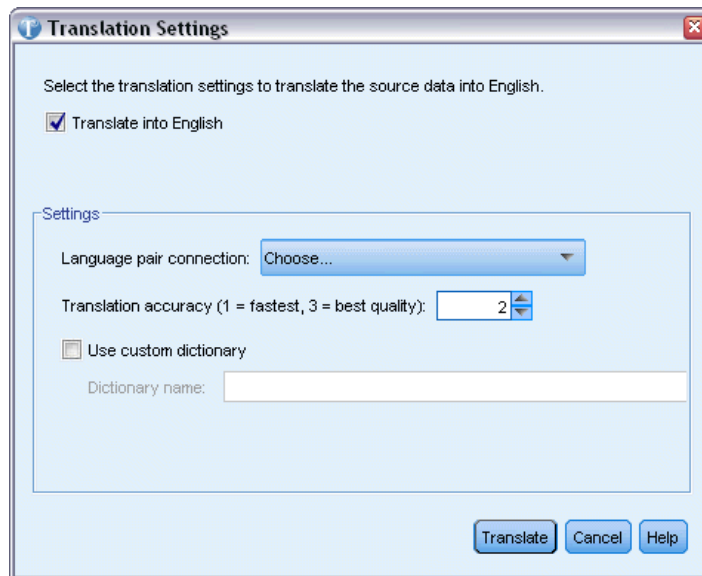
- ▶ From the menus choose File > Update Data. The data are read again to take into account your new changes.
- ▶ If a translation in English was performed before, the Translation Settings dialog appears so that you can choose the language pair and translate again. For more information, see the topic “Translating into English” on p. 71.

Translating into English

You can refresh a translation whenever you want. After translating, you will need to extract again since your translation results will be out of sync with your new translation.

Note: If you want to translate new data, you can do so directly in the New Project wizard when you create a new project. For more information, see the topic “Translating into English” on p. 34.

Figure 4-19
Translation Settings



To Translate Into English

- ▶ From the menus choose Tools > Translation Settings. The Translation Settings dialog appears.
- ▶ To translate the text data from a licensed language into English, select the Translate into English checkbox.
- ▶ From the Language Pair Connection list, select the connection for the Language Weaver language pair you want to use. If you have Language Weaver configured on your local machine, those language pairs will automatically appear in this list. You can add or test network (WAN) or online services (HTTP) connections in the Translation tab of the Options dialog. For more information, see the topic “Options: Translation Tab” in Chapter 2 on p. 21.
- ▶ Specify the desired Translation accuracy. Choose a value of 1 to 3 indicating the level of speed versus accuracy you want. A lower value produces faster translation results but with diminished accuracy. A higher value produces results with greater accuracy but increased processing time. To optimize time, we recommend beginning with a lower level and increasing it only if you feel you need more accuracy after reviewing the results.
- ▶ If you have previously created custom dictionaries, held by Language Weaver, you can use them in connection with the translation. To choose a custom dictionary, select the Use custom dictionary checkbox and enter the Dictionary name. To use more than one dictionary, separate the names with a comma.
- ▶ Click Translate to begin the translation process. The translation progress dialog appears.

Sharing Projects

You can share your projects with other users or if you want to work on a project on another machine.

To share a project:

- ▶ From the menus choose File > Save Project. The project is saved.
- ▶ Send the project file to another machine or person. This project file contains a reference to the data file you originally imported. If you want the other user to be able to use the same source data for this project, you must also provide them with the original data file and inform them of the path in which they should copy that data file so that IBM® SPSS® Text Analytics for Surveys can find the data file when the project file is opened.
- ▶ When the other user opens the project file in SPSS Text Analytics for Surveys, he or she can choose between using the local libraries contained in the project file or using public versions of these libraries they already had. Generally, to ensure the same results, the local versions should be used.
- ▶ If SPSS Text Analytics for Surveys cannot locate the data file, a message appears and warns the user that the data must be reimported. For more information, see the topic “Changing Data Sources” on p. 61.

Flagging Responses

To help you monitor your progress as you analyze your survey, you can mark responses using flags in the Data pane. There are many reasons why you might want to mark a response, including:

- To mark off the responses that you have manually reviewed so that you know where to pick up later
- To mark off a response that you are unsure about how to handle
- To mark and export the flags into another program

Once you mark a response with a flag, you can continue to work with these responses. They are purely for your own record-keeping. You can choose among the following flags:

Table 4-1
Flag descriptions



Flag	Description
	Complete flag to denote responses that you deem finished.
	Important flag to denote responses that you deem important.

Figure 4-20
Response flags in the Data pane

	Id	Response	Categories
1	1	little, light	Pos: Size-Weight
2	3	cost and size	Pos: Pricing and Billing Pos: Size-Weight
3	9	Small , great sound , capacity .	Pos: Features-Design Pos: Size-Weight Pos: Storage
4	20	lightweight	Pos: Size-Weight
5	25	very small and holds lots of songs	Pos: Size-Weight
6	35	small , easy to sync	Pos: Size-Weight Pos: Usability
7	46	Ability to carry large amounts of music in a small , lightweight device .	Pos: Features-Design Pos: Size-Weight
8	52	i have a Product A . I like the small size and good sound .	Pos: Features-Design Pos: Size-Weight
9	54	The ability to take my CD collection and have it in one simple , small , portable device .	Pos: Usability Pos: Size-Weight

To mark a response with a flag:

- ▶ From within the Data pane, select the response that you want to mark.
- ▶ From the menus choose Edit > Mark Responses With and then select the type of flag that you want to use (Important Flag or Complete Flag). The selected flag is assigned. If the Flag column in the Data pane is not visible, it appears. The status bar is updated with the number of flagged responses.

To clear flags:

- ▶ From within the Data pane, right-click on the responses for which you want to remove a flag.
- ▶ From the context menu choose Mark Responses With > Clear Flags. The selected flags are removed.

Project Status Bar

Depending on the window or view you are working in, different status bars exist. By default, a status bar is displayed whenever you have an open project. This status bar provides summary information about the project and the elements it contains. You can also turn the status bar on and off when desired.

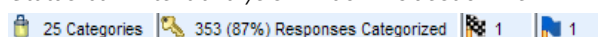
To disable or enable the status bar in either window:

- ▶ From the menus choose View > Status Bar.

Text Analysis Window

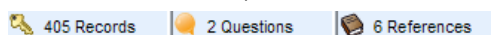
This status bar provides summary information about questions and responses in the project. Depending on where you are in the text analysis window, the information in the status bar changes. You can also see the number of responses that have been marked as important or complete.

Figure 4-21
Status bar in text analysis window: Question view



When you are in the Question view, you see the number of categories for that question and the percentage-of-response categorization. When you are in the Entire Project view, you see information for the entire project.

Figure 4-22
Status bar in text analysis window: Entire Project view



The following table describes each element in the status bar.

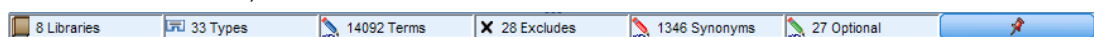
Table 4-2
Text analysis window: Status bar description

Element	Description
Records	The number of records in your data.
Questions	The number of questions in your project.
Reference	The number of imported reference variables. Reference variables are additional variables that are imported for reference purposes but are not analyzed.
Categories	The number of categories for a given question. If empty categories exist, then the number of empty categories appears in parentheses.
Categorized	The number of responses for the question followed by the categorization percentage in parentheses.

Resource Editor Window

This status bar provides information about the linguistic resources for the project. The forced terms area in the bar is actionable, meaning that you can click on it to take action. When working with libraries, you can disable elements within the libraries to exclude them from processing. For more information, see the topic “Disabling Local Libraries” in Chapter 9 on p. 200. If the project contains disabled elements, two number counts appear in the status bar—the first is the number of elements present, and the second is the number of enabled elements. For example, if your status bar shows 5(2) Libraries, this means that there are five libraries in your project but that only two are enabled.

Figure 4-23
Status bar in Dictionary Editor window



The following table describes each element in the status bar.

Table 4-3
Resource Editor window: Status bar description

Element	Description
Library	The number of libraries in the project.
Type	The number of types in the entire project.
Term	The number of terms in all libraries. If a term is in the Exclude list, it is still included in the count. Note that if a type is disabled, all terms in that type are also disabled.
Exclude	The number of excluded items in all libraries in the project.

Element	Description
Synonym	The number of synonym targets in all libraries in the project.
Optional	The number of defined optional elements in all libraries in the project. Please note that the counts include each delimited entry in a cell individually.
Forced terms	A button that is enabled whenever there are forced terms in the libraries of your project. Clicking this button displays the Edit Forced Terms dialog box. For more information, see the topic “Forcing Terms” in Chapter 10 on p. 214.

Extracting Data

When you create a project through the New Project wizard, the default choice is to perform an extraction automatically for the first question. If you want to refresh an extraction or extract for a new question, you can do so manually (Tools > Extract) or choose to extract when you start building categories. The end result of this extraction is a set of concepts, types, and patterns. You can view and work with these results in the Extraction Results pane.

If an extraction was not performed when you created your project or if you chose not to save extraction results, then you can navigate to the question you want to start working (View > Question > “Question”) and then extract.

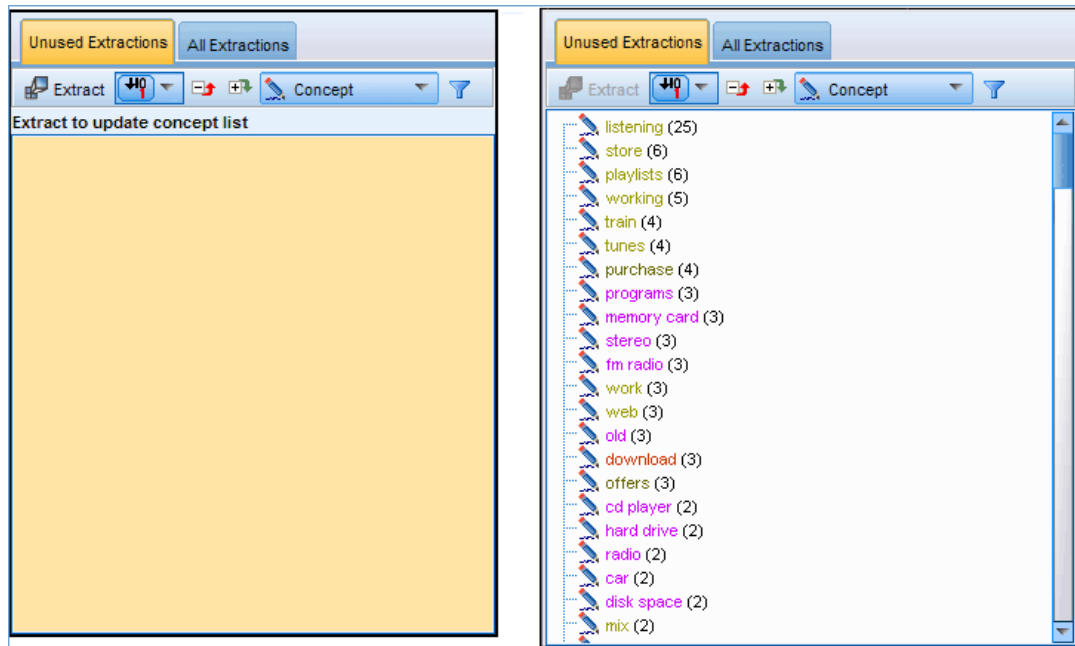
After extracting, you should review the results and make any changes that you find necessary. For more information, see the topic “Refining Extraction Results” on p. 84. You can then reextract to see the new results. When manually coding responses, two individuals might group responses slightly differently. However, accuracy and continuity are extremely important in categorizing survey responses.

The power of IBM® SPSS® Text Analytics for Surveys lies in its ability to provide the consistent reapplication of your category definitions. By fine-tuning your extraction results from the start, you can be assured that each time you reextract, you will get identical results in your category definitions, well-adapted to the context of the data. In this way, responses will be assigned to your category definitions in a more accurate, repeatable manner.

Extracted Results: Concepts, Types, and Patterns

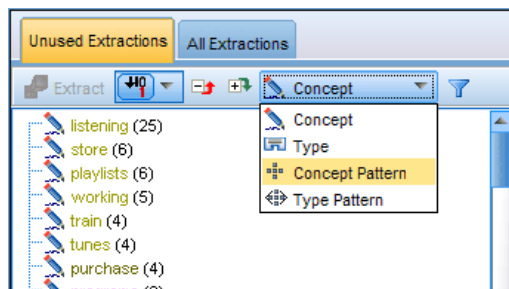
After you create a project, the window automatically displays the first open-ended question that you imported. The Extraction Results pane is located in the lower left corner of the Question view. This view is accessed from the View menu (View > Question > “Question”). If no extraction results exist, you must extract to begin working. For more information, see the topic “Extracting Data” on p. 81.

Figure 5-1
Extraction results pane before and after extraction



If the Extraction Results pane is empty or out of date, it is colored in yellow. Click the Extract button to launch the extraction process. After you extract, you can look at the results by selecting what you want to see from the drop-down list.

Figure 5-2
Extraction results pane drop-down list



The concepts, types, and TLA patterns that are extracted are collectively referred to as **extraction results**, and they serve as the descriptors, or building blocks, for your categories. You can also use concepts, types, and patterns in your category rules. Additionally, the automatic techniques use concepts and types to build the categories.

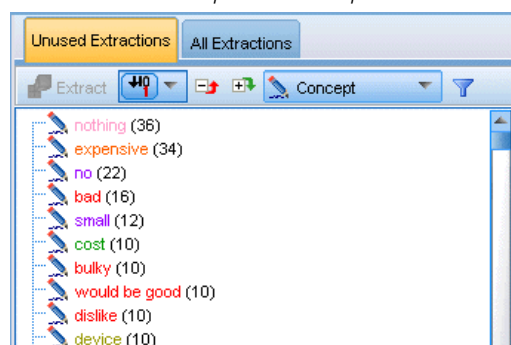
Text analysis is an iterative process in which extraction results are reviewed according to the context of the text data, fine-tuned to produce new results, and then reevaluated. After extracting, you should review the results and make any changes that you find necessary by modifying the linguistic resources. You can fine-tune the resources, in part, directly from the Extraction Results pane or Data pane through context menus. For more information, see the topic “Refining Extraction Results” on p. 84. You can also do so directly in the Resource Editor view. For more information, see the topic “The Resource Editor Window” in Chapter 2 on p. 14.

After fine-tuning, you can then reextract to see the new results. By fine-tuning your extraction results from the start, you can be assured that each time you reextract, you will get identical results in your category definitions, perfectly adapted to the context of the data. In this way, records will be assigned to your category definitions in a more accurate, repeatable manner.

Concepts

During the extraction process, the text data is scanned and analyzed in order to identify interesting or relevant single words (such as `election` or `peace`) and word phrases (such as `presidential election`, `election of the president`, or `peace treaties`) in the text. These words and phrases are collectively referred to as *terms*. Using the linguistic resources, the relevant terms are extracted and then similar terms are grouped together under a lead term called a **concept**.

Figure 5-3
Extraction Results pane: Concept view



By default, the concepts are shown in lowercase and sorted in descending order according to the number of records in which the concept is found. When concepts are extracted, they are assigned a type to help group similar concepts. They are color coded according to this type. Colors are defined in the type properties within the Resource Editor. For more information, see the topic “Type Dictionaries” in Chapter 10 on p. 207.

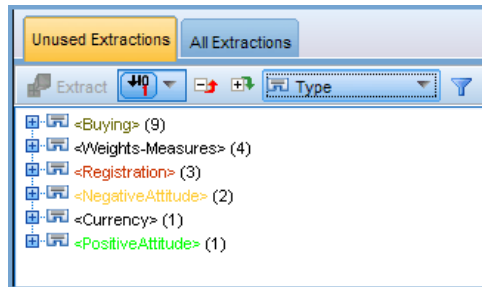
Whenever a concept, type, or pattern is being used in a category definition, it appears in italics in the table.

Types

Types are semantic groupings of concepts. When concepts are extracted, they are assigned a type to help group similar concepts. Several built-in types are delivered with IBM® SPSS® Text Analytics for Surveys, such as `<Location>`, `<Organization>`, `<Person>`, `<Positive>`, `<Negative>` and so on. For example, the `<Location>` type groups geographical keywords and places. This type would be assigned to concepts such as `chicago`, `paris`, and `tokyo`. Concepts that are not found in any type dictionary but are extracted from the text are automatically typed as `<Unknown>`. For more information, see the topic “Built-in Types” in Chapter 10 on p. 208.

When you select the Type view, the extracted types appear by default in descending order by frequency. When the tree is expanded, you see the concepts that were extracted for that type. You can also see that types are color coded to help distinguish them. Colors are part of the type properties. For more information, see the topic “Creating Types” in Chapter 10 on p. 209. You can also create your own types.

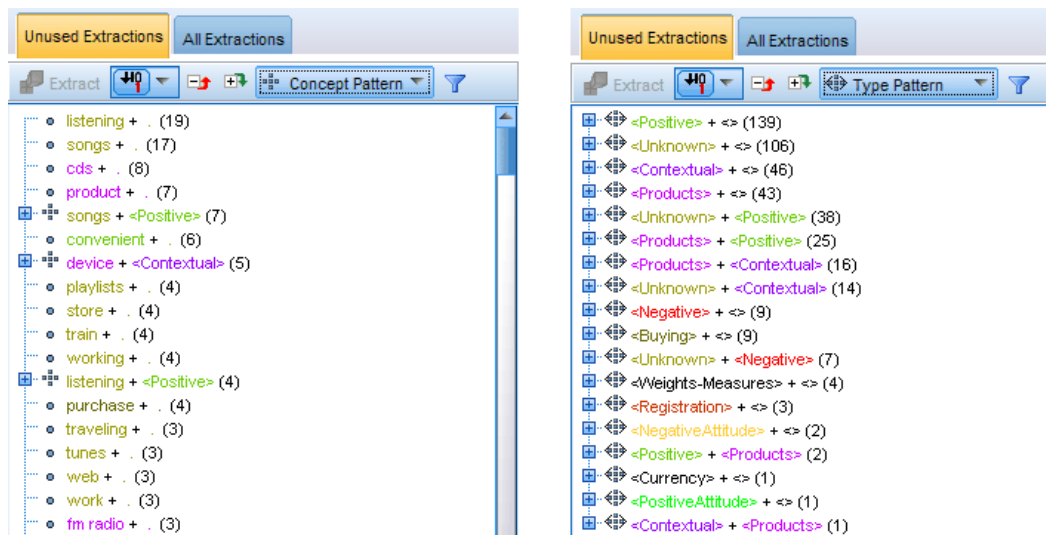
Figure 5-4
Extraction Results pane: Type view



Patterns

Patterns are made up of two parts: a combination of concepts and built-in types representing qualifiers and adjectives. Patterns are most useful when you are attempting to discover opinions about a particular subject. Extracting your competitor's product name may not be interesting enough to you. In this case, you can look at the extracted patterns to see if you can find examples where respondents found the product to be good, bad, or expensive. There are two different pattern views: Concept Patterns and Type Patterns.

Figure 5-5
Extraction Results pane: Concept Pattern view and Type Pattern view



Concept Pattern. In this view, the top level of the tree in the Extraction Results pane displays patterns with the following structure: `concept1 + <Type1>` for concept patterns, such as `text analysis + <Positive>` or `cost + <Negative>`. When the tree is expanded, you can see the exact patterns, such as `text analysis + powerful` or `cost + expensive`.

Patterns can also be significant when they exist without a second part of the pattern. For example, you may be interested in finding occurrences where the respondent did not express a negative or positive opinion about the subject. In this case, this is represented as `concept1 + .`, where `.` designates a null qualifier. For example, if a respondent answered, “*Cost, store location*” when answering the question, “*What factors influence your decision to choose a music player?*”,

the extraction could produce `cost + .` and `store_location + .` as null patterns. When patterns are displayed, colors are attributed to each element in the pattern depending on their type.

Type Pattern. In this view, the top level of the tree in the Extraction Results pane displays patterns with the following structure: `<Type> + <Type>`, such as `<Budget> + <Positive>`. If you expand the tree further, you will see relationships as described and presented in the Concept Pattern view. When patterns are displayed, colors are attributed to each element in the pattern depending on their type.

Unused Extractions and All Extractions Tabs

The Extraction Results pane presents the output from the extraction process. As you begin creating categories, some of the extraction results (concepts, types, and patterns) will become part of the category descriptors. For this reason, SPSS Text Analytics for Surveys presents this information in two ways using tabs. You can switch back and forth between viewing those elements that are already used in category definitions or the entire set of extracted concepts. You can do this by clicking the Unused Extractions and All Extractions tabs. The Unused Extractions tab displays all elements that are not currently part of a category descriptor. The All Extractions tab displays all extracted items with the used items appearing in italics.

Extracting Data

Whenever an extraction is needed, the Extraction Results pane becomes yellow in color and the message Press Extract Button to Extract Concepts appears below the toolbar in this pane.

You may need to extract if you do not have any extraction results yet, have made changes to the linguistic resources and need to update the extraction results, or have reopened a project in which you did not save the extraction results (Tools > Options).

Note: If you change the source node for your stream after extraction results have been cached with the Use session work... option, you will need to run a new extraction once the interactive workbench session is launched if you want to get updated extraction results.

When you run an extraction, a progress indicator appears to provide feedback on the status of the extraction. During this time, the extraction engine reads through all of the text data and identifies the relevant terms and patterns and extracts them and assigns them to a type. Then, the engine attempts groups synonyms terms under one lead term, called a concept. When the process is complete, the resulting concepts, types, and patterns appear in the Extraction Results pane. You can begin working with and reviewing the results.

Note: There is a relationship between the size of your dataset and the time it takes to complete the extraction process. See the installation instructions for performance statistics and recommendations.

To Extract Data

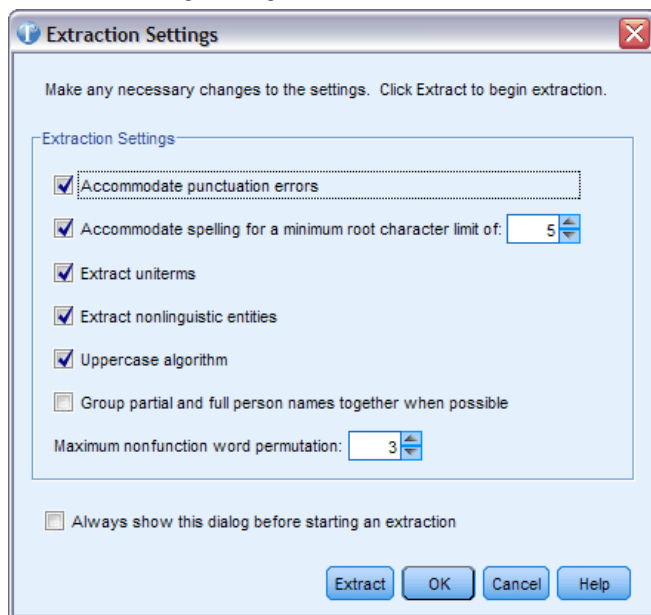
- ▶ From the menus, choose Tools > Extract. Alternatively, click the Extract toolbar button.
- ▶ If you chose to always display the Extraction Settings dialog, it appears so that you can make any changes. See further in this topic for descriptors of each settings.

- ▶ Click Extract to begin the extraction process. Once the extraction begins, the progress dialog box opens. After extraction, the results appear in the Extraction Results pane. By default, the concepts are shown in lowercase and sorted in descending order according to the number of records in which the concept is found.

You can review the results using the toolbar options to sort the results differently, to filter the results, or to switch to a different view (concepts, patterns, or types). You can also refine your extraction results by working with the linguistic resources. For more information, see the topic “Refining Extraction Results” on p. 84.

The Extraction Settings dialog box contains some basic extraction options.

Figure 5-6
Extraction Settings dialog box



Accommodate punctuation errors. This option temporarily normalizes text containing punctuation errors (for example, improper usage) during extraction to improve the extractability of concepts. This option is extremely useful when text is short and of poor quality (as, for example, in open-ended survey responses, e-mail, and CRM data), or when the text contains many abbreviations.

Accommodate spelling errors for a minimum root character limit of [n]. This option applies a fuzzy grouping technique that helps group commonly misspelled words or closely spelled words under one concept. The fuzzy grouping algorithm temporarily strips all vowels (except the first one) and strips double/triple consonants from extracted words and then compares them to see if they are the same so that `modeling` and `modelling` would be grouped together. However, if each term is assigned to a different type, excluding the <Unknown> type, the fuzzy grouping technique will not be applied.

You can also define the minimum number of *root* characters required before fuzzy grouping is used. The number of root characters in a term is calculated by totaling all of the characters and subtracting any characters that form inflection suffixes and, in the case of compound-word

terms, determiners and prepositions. For example, the term *exercises* would be counted as 8 root characters in the form “exercise,” since the letter *s* at the end of the word is an inflection (plural form). Similarly, *apple sauce* counts as 10 root characters (“apple sauce”) and *manufacturing of cars* counts as 16 root characters (“manufacturing car”). This method of counting is only used to check whether the fuzzy grouping should be applied but does not influence how the words are matched.

Note: If you find that certain words are later grouped incorrectly, you can exclude word pairs from this technique by explicitly declaring them in the Fuzzy Grouping: Exceptions section in the Advanced Resources tab. For more information, see the topic “Fuzzy Grouping” in Chapter 11 on p. 227.

Extract uniterms. This option extracts single words (uniterms) as long as the word is not already part of a compound word and if it is either a noun or an unrecognized part of speech.

Extract nonlinguistic entities. This option extracts nonlinguistic entities, such as phone numbers, social security numbers, times, dates, currencies, digits, percentages, e-mail addresses, and HTTP addresses. You can include or exclude certain types of nonlinguistic entities in the Nonlinguistic Entities: Configuration section of the Advanced Resources tab. By disabling any unnecessary entities, the extraction engine won’t waste processing time. For more information, see the topic “Configuration” in Chapter 11 on p. 232.

Uppercase algorithm. This option extracts simple and compound terms that are not in the built-in dictionaries as long as the first letter of the term is in uppercase. This option offers a good way to extract most proper nouns.

Group partial and full person names together when possible. This option groups names that appear differently in the text together. This feature is helpful since names are often referred to in their full form at the beginning of the text and then only by a shorter version. This option attempts to match any uniterm with the <Unknown> type to the last word of any of the compound terms that is typed as <Person>. For example, if *doe* is found and initially typed as <Unknown>, the extraction engine checks to see if any compound terms in the <Person> type include *doe* as the last word, such as *john doe*. This option does not apply to first names since most are never extracted as uniterms.

Maximum nonfunction word permutation. This option specifies the maximum number of nonfunction words that can be present when applying the permutation technique. This permutation technique groups similar phrases that differ from each other only by the nonfunction words (for example, *of* and *the*) contained, regardless of inflection. For example, let’s say that you set this value to at most two words and both *company officials* and *officials of the company* were extracted. In this case, both extracted terms would be grouped together in the final concept list since both terms are deemed to be the same when *of the* is ignored.

Always show this dialog before starting an extraction. Specify whether you want to see the Extraction Settings dialog each time you extract, if you never want to see it unless you go to the Tools menu, or whether you want to be asked each time you extract if you want to edit any extraction settings.

Saving Extraction Results

Whenever you extract, the results appear in the Extraction Results pane and can be used to categorize your responses. During an IBM® SPSS® Text Analytics for Surveys session, these extraction results are held in memory so that you can work with them. By default, extraction results are saved in your projects. Whether or not you save them when closing the project is a global setting that you can change at any time in the Options dialog box (Tools > Options). For more information, see the topic “Options: System Tab” in Chapter 2 on p. 17.

As a security measure, these extraction results are encrypted during the save process and placed in the database. This procedure makes it difficult for someone to come across any data in the database. Furthermore, extraction results are never presented in SPSS Text Analytics for Surveys until the data source has been located by the application. Therefore, if the data are password-protected, a user must enter the user name and password for this data source before the extraction results appear on the screen.

Refining Extraction Results

Extraction is an iterative process whereby you can extract, review the results, make changes to them, and then reextract to update the results. Since accuracy and continuity are essential to successful text mining and categorization, fine-tuning your extraction results from the start ensures that each time you reextract, you will get precisely the same results in your category definitions. In this way, records will be assigned to your categories in a more accurate, repeatable manner.

The extraction results serve as the building blocks for categories. When you create categories using these extraction results, records are automatically assigned to categories if they contain text that matches one or more category descriptors. Although you can begin categorizing before making any refinements to the linguistic resources, it is useful to review your extraction results at least once before beginning.

As you review your results, you may find elements that you want the extraction engine to handle differently. Consider the following examples:

- **Unrecognized synonyms.** Suppose you find several concepts you consider to be synonymous, such as *smart*, *intelligent*, *bright*, and *knowledgeable*, and they all appear as individual concepts in the extraction results. You could create a synonym definition in which *intelligent*, *bright*, and *knowledgeable* are all grouped under the target concept *smart*. Doing so would group all of these together with *smart*, and the global frequency count would be higher as well. For more information, see the topic “Adding Synonyms” on p. 85.
- **Mistyped concepts.** Suppose that the concepts in your extraction results appear in one type and you would like them to be assigned to another. In another example, imagine that you find 15 vegetable concepts in your extraction results and you want them all to be added to a new type called `<Vegetable>`. Concepts that are not found in any type dictionary but are extracted from the text are automatically typed as `<Unknown>`. You can add concepts to types. For more information, see the topic “Adding Concepts to Types” on p. 87.
- **Insignificant concepts.** Suppose that you find a concept that was extracted and has a very high frequency count—that is, it is found in many records. However, you consider this concept to be insignificant to your analysis. You can exclude it from extraction. For more information, see the topic “Excluding Concepts from Extraction” on p. 89.

- **Incorrect matches.** Suppose that in reviewing the records that contain a certain concept, you discover that two words were incorrectly grouped together, such as `faculty` and `facility`. This match may be due to an internal algorithm, referred to as fuzzy grouping, that temporarily ignores double or triple consonants and vowels in order to group common misspellings. You can add these words to a list of word pairs that should not be grouped. For more information, see the topic “Fuzzy Grouping” in Chapter 11 on p. 227.
- **Unextracted concepts.** Suppose that you expect to find certain concepts extracted but notice that a few words or phrases were not extracted when you review the record text. Often these words are verbs or adjectives that you are not interested in. However, sometimes you do want to use a word or phrase that was not extracted as part of a category definition. To extract the concept, you can force a term into a type dictionary. For more information, see the topic “Forcing Words into Extraction ” on p. 90.

Many of these changes can be performed directly from the Extraction Results pane or Data pane by selecting one or more elements and right-clicking your mouse to access the context menus.

After making your changes, the pane background color changes to show that you need to reextract to view your changes. For more information, see the topic “Extracting Data” on p. 81. If you are working with larger datasets, it may be more efficient to reextract after making several changes rather than after each change.

Note: You can view the entire set of editable linguistic resources used to produce the extraction results in the Resource Editor view (View > Resource Editor). These resources appear in the form of libraries and dictionaries in this view. You can customize the concepts and types directly within the libraries and dictionaries. For more information, see the topic “Working with Libraries” in Chapter 9 on p. 195.

Adding Synonyms

Synonyms associate two or more words that have the same meaning. Synonyms are often also used to group terms with their abbreviations or to group commonly misspelled words with the correct spelling. By using synonyms, the frequency for the target concept is greater, which makes it far easier to discover similar information that is presented in different ways in your text data.

The linguistic resource templates and libraries delivered with the product contain many predefined synonyms. However, if you discover unrecognized synonyms, you can define them so that they will be recognized the next time you extract.

The first step is to decide what the target, or lead, concept will be. The **target concept** is the word or phrase under which you want to group all synonym terms in the final results. During extraction, the synonyms are grouped under this target concept. The second step is to identify all of the synonyms for this concept. The target concept is substituted for all synonyms in the final extraction. A term must be extracted to be a synonym. However, the target concept does not need to be extracted for the substitution to occur. For example, if you want `intelligent` to be replaced by `smart`, then `intelligent` is the synonym and `smart` is the target concept.

If you create a new synonym definition, a new target concept is added to the dictionary. You must then add synonyms to that target concept. Whenever you create or edit synonyms, these changes are recorded in synonym dictionaries in the Resource Editor. If you want to view the entire contents of these synonym dictionaries or if you want to make a substantial number of

changes, you may prefer to work directly in the Resource Editor. For more information, see the topic “Substitution/Synonym Dictionaries” in Chapter 10 on p. 217.

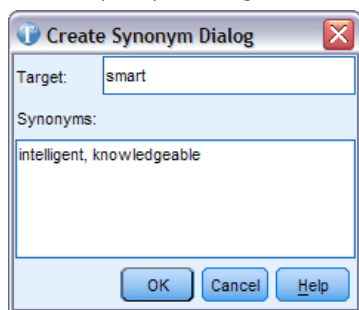
Any new synonyms will automatically be stored in the first library listed in the library tree in the Resource Editor view—by default, this is the *Local Library*.

Note: If you look for a synonym definition and cannot find it through the context menus or directly in the Resource Editor, a match may have resulted from an internal fuzzy grouping technique. For more information, see the topic “Fuzzy Grouping” in Chapter 11 on p. 227.

To Create a New Synonym

- ▶ In either the Extraction Results pane or Data pane, select the concept(s) for which you want to create a new synonym.
- ▶ From the menus, choose Edit > Add to Synonym > New. The Create Synonym dialog box opens.

Figure 5-7
Create Synonym dialog box

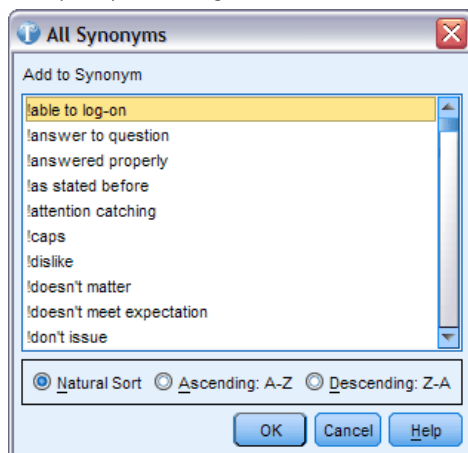


- ▶ Enter a target concept in the Target text box. This is the concept under which all of the synonyms will be grouped.
- ▶ If you want to add more synonyms, enter them in the Synonyms list box. Use the global separator to separate each synonym term. For more information, see the topic “Options: System Tab” in Chapter 2 on p. 17.
- ▶ Click OK to apply your changes. The dialog box closes and the Extraction Results pane background color changes, indicating that you need to reextract to see your changes. If you have several changes, make them before you reextract.

To Add to a Synonym

- ▶ In either the Extraction Results pane or Data pane, select the concept(s) that you want to add to an existing synonym definition.
- ▶ From the menus, choose Edit > Add to Synonym > . The menu displays a set of the synonyms with the most recently created at the top of the list. Select the name of the synonym to which you want to add the selected concept(s). If you see the synonym that you are looking for, select it, and the concept(s) selected are added to that synonym definition. If you do not see it, select More to display the All Synonyms dialog box.

Figure 5-8
All Synonyms dialog box



- In the All Synonyms dialog box, you can sort the list by natural sort order (order of creation) or in ascending or descending order. Select the name of the synonym to which you want to add the selected concept(s) and click OK. The dialog box closes, and the concepts are added to the synonym definition.

Adding Concepts to Types

Whenever an extraction is run, the extracted concepts are assigned to types in an effort to group terms that have something in common. IBM® SPSS® Text Analytics for Surveys is delivered with many built-in types. For more information, see the topic “Built-in Types” in Chapter 10 on p. 208. Concepts that are not found in any type dictionary but are extracted from the text are automatically typed as <Unknown>.

When reviewing your results, you may find some concepts that appear in one type that you want assigned to another, or you may find that a group of words really belongs in a new type by itself. In these cases, you would want to reassign the concepts to another type or create a new type altogether.

For example, suppose that you are working with survey data relating to automobiles and you are interested in categorizing by focusing on different areas of the vehicles. You could create a type called <Dashboard> to group all of the concepts relating to gauges and knobs found on the dashboard of the vehicles. Then you could assign concepts such as gas gauge, heater, radio, and odometer to that new type.

In another example, suppose that you are working with survey data relating to universities and colleges and the extraction typed Johns Hopkins (the university) as a <Person> type rather than as an <Organization> type. In this case, you could add this concept to the <Organization> type.

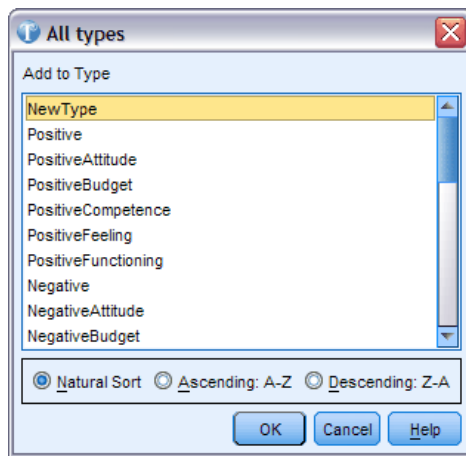
Whenever you create a type or add concepts to a type’s term list, these changes are recorded in type dictionaries within your linguistic resource libraries in the Resource Editor. If you want to view the contents of these libraries or make a substantial number of changes, you may prefer to work directly in the Resource Editor. For more information, see the topic “Adding Terms” in Chapter 10 on p. 210.

To Add a Concept to a Type

- ▶ In either the Extraction Results pane or Data pane, select the concept(s) that you want to add to an existing type.
- ▶ Right-click to open the context menu.
- ▶ From the menus, choose Edit > Add to Type >. The menu displays a set of the types with the most recently created at the top of the list. Select the type name to which you want to add the selected concept(s). If you see the type name that you are looking for, select it, and the concept(s) selected are added to that type. If you do not see it, select More to display the All Types dialog box.

Figure 5-9

All Types dialog box

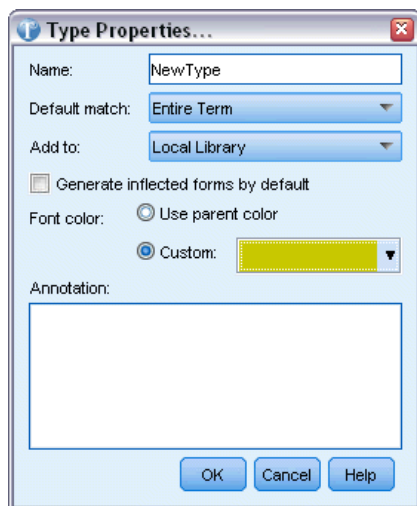


- ▶ In the All Types dialog box, you can sort the list by natural sort (order of creation) or in ascending or descending order. Select the name of the type to which you want to add the selected concept(s) and click OK. The dialog box closes, and they are added as terms to the type.

To Create a New Type

- ▶ In either the Extraction Results pane or Data pane, select the concepts for which you want to create a new type.
- ▶ From the menus, choose Edit > Add to Type > New. The Type Properties dialog box opens.

Figure 5-10
Type Properties dialog box



- ▶ Enter a new name for this type in the Name text box and make any changes to the other fields. For more information, see the topic “Creating Types” in Chapter 10 on p. 209.
- ▶ Click OK to apply your changes. The dialog box closes and the Extraction Results pane background color changes, indicating that you need to reextract to see your changes. If you have several changes, make them before you reextract.

Excluding Concepts from Extraction

When reviewing your results, you may occasionally find concepts that you did not want extracted or used by any automated category building techniques. In some cases, these concepts have a very high frequency count and are completely insignificant to your analysis. In this case, you can mark a concept to be excluded from the final extraction. Typically, the concepts you add to this list are fill-in words or phrases used in the text for continuity but that do not add anything important and may clutter the extraction results. By adding concepts to the exclude dictionary, you can make sure that they are never extracted.

By excluding concepts, all variations of the excluded concept disappear from your extraction results the next time that you extract. If this concept already appears as a descriptor in a category, it will remain in the category with a zero count after reextraction.

When you exclude, these changes are recorded in an exclude dictionary in the Resource Editor. If you want to view all of the exclude definitions and edit them directly, you may prefer to work directly in the Resource Editor. For more information, see the topic “Exclude Dictionaries” in Chapter 10 on p. 222.

To Exclude Concepts

- ▶ In either the Extraction Results pane or Data pane, select the concept(s) that you want to exclude from the extraction.
- ▶ Right-click to open the context menu.

- ▶ Select Exclude from Extraction. The concept is added to the exclude dictionary in the Resource Editor and the Extraction Results pane background color changes, indicating that you need to reextract to see your changes. If you have several changes, make them before you reextract.

Note. Any words that you exclude will automatically be stored in the first library listed in the library tree in the Resource Editor—by default, this is the *Local Library*.

Forcing Words into Extraction

When reviewing the text data in the Data pane after extraction, you may discover that some words or phrases were not extracted. Often, these words are verbs or adjectives that you are not interested in. However, sometimes you do want to use a word or phrase that was not extracted as part of a category definition.

If you would like to have these words and phrases extracted, you have two options:

- Force a term into a type library. For more information, see the topic “Forcing Terms” in Chapter 10 on p. 214.
- Add words directly to an existing category definition. This is generally used if the first option did not provide the expected results. For more information, see the topic “Text Matching in Categories” in Chapter 6 on p. 154.

Important! Marking a term in a dictionary as forced is not foolproof. By this, we mean that even though you have explicitly added a term to a dictionary, there are times when it may not be present in the Extraction Results pane after you have reextracted or it does appear but not exactly as you have declared it. Although this occurrence is rare, it can happen when a word or phrase was already extracted as part of a longer phrase. To prevent this, apply the Entire (no compounds) match option to this term in the type dictionary. For more information, see the topic “Adding Terms” in Chapter 10 on p. 210.

Categorizing Text Data

In IBM® SPSS® Text Analytics for Surveys, you can create **categories** that represent, in essence, higher-level concepts, or topics, that will capture the key ideas, knowledge, and attitudes expressed in the text.

Categories can also have a hierarchical structure, meaning they can contain subcategories and those subcategories can also have subcategories of their own and so on. You can import predefined category structures, formerly called code frames, with hierarchical categories as well as build these hierarchical categories inside the product.

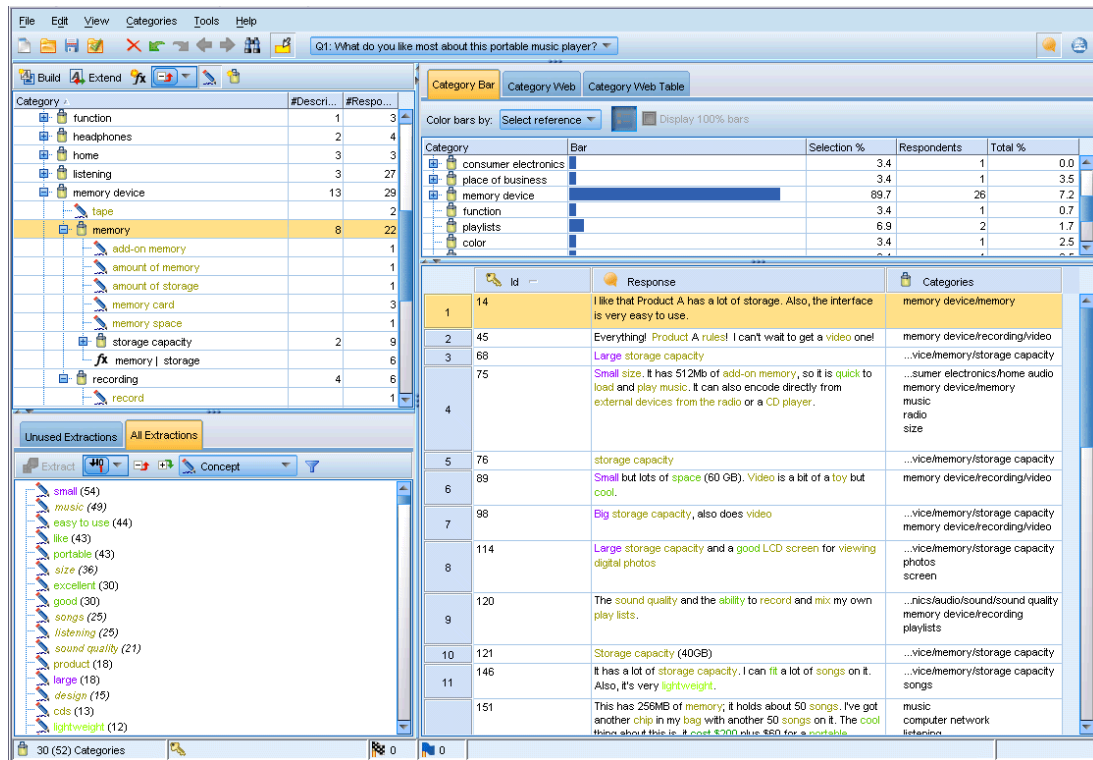
In effect, hierarchical categories enable you to build a tree structure with one or more subcategories to group items such as different concept or topic areas more accurately. A simple example can be related to leisure activities; answering a question such as *What activity would you like to do if you had more time?* you may have top categories such as *sports, art and craft, fishing*, and so on; down a level, below *sports*, you may have subcategories to see if this is *ball games, water-related*, and so on.

Categories are made up of a set of descriptors, such as *concepts, types, patterns* and *category rules*. Together, these descriptors are used to identify whether or not a record belongs to a given category. The text within a record can be scanned to see whether any text matches a descriptor. If a match is found, the record is assigned to that category. This process is called **categorization**.

You can work with, build, and visually explore your categories using the data presented in the four panes, each of which can be hidden or shown by selecting its name from the View menu.

- **Categories pane.** Build and manage your categories in this pane. For more information, see the topic “The Categories Pane” on p. 92.
- **Extraction Results pane.** Explore and work with the extracted concepts and types in this pane. For more information, see the topic “Extracted Results: Concepts, Types, and Patterns” in Chapter 5 on p. 77.
- **Visualization pane.** Visually explore your categories and how they interact in this pane. For more information, see the topic “Visualizing Graphs” in Chapter 7 on p. 159.
- **Data pane.** Explore and review the text contained within records that correspond to selections in this pane. For more information, see the topic “The Data Pane” on p. 95.

Figure 6-1
Question view



While you might start with a set of categories from a text analysis package (TAP) or import from a predefined category file, you might also need to create your own. Categories can be created automatically using the product's robust set of automated techniques, which use extraction results (concepts, types, and patterns) to generate categories and their descriptors. Categories can also be created manually using additional insight you may have regarding the data. You can create category definitions manually by dragging and dropping extraction results into the categories. You can enrich these categories or any empty category by adding category rules to a category, using your own predefined categories, adding a word or phrase that was never extracted (called **text matching**), by forcing responses directly into a category, or a combination.

Each of the techniques and methods is well suited for certain types of data and situations, but often it will be helpful to combine techniques in the same analysis to capture the full range of records. And in the course of categorization, you may see other changes to make to the linguistic resources.

The Categories Pane

The Categories pane is the area in which you can build and manage your categories. This pane is located in the upper left corner of the Question view and is accessible from the View menu (View > Question > "Your_Question"). After extracting the concepts and types from your text data, you can begin building categories automatically using techniques such as concept inclusion, co-occurrence, and so on or manually. For more information, see the topic "Building Categories" on p. 105.

Figure 6-2
Categories pane without categories and with categories

Category	#Descriptors	#Responses
All Records	-	405
Uncategorized	-	405
No concepts extracted	-	405

Category	#Descriptors	#Responses
All Records	-	405
Uncategorized	-	145
No concepts extracted	-	4
music	25	73
child's music		1
music catalogue		1
play music		2
music to listen		1
music choice		1
amounts of music		1
library of music		1
collection of music		1
bank of music		1
digital music hands		1
share music		1

Each time a category is created or updated, the records are scanned automatically to see whether any text matches a descriptor in a given category. If a match is found, the record is assigned to that category. The end result is that most, if not all, of the records are assigned to categories based on the descriptors in the categories.

Category Tree Table

The tree table in this pane presents the set of categories, subcategories, and descriptors. The tree also has several columns presenting information for each tree item. The following columns may be available for display:

- **Code.** Lists the code value for each category. This column is hidden by default. You can display this column by right-clicking in the tree table and selecting Show > Category Code.
- **Category.** Contains the category tree showing the name of the category and subcategories. Additionally if the descriptors toolbar icon is clicked, the set of descriptors will also be displayed.
- **Descriptors.** Provides the number of descriptors that make up its definition. This count does not include the number of descriptors in the subcategories. No count is given when a descriptor name is shown in the Categories column. You can display this column by right-clicking in the tree table and selecting Show > All Descriptors.
- **Docs.** After scoring, this column provides the number of records that are categorized into a category and all of its subcategories. So if 5 records match your top category based on its descriptors, and 7 different records match a subcategory based on its descriptors, the total doc count for the top category is a sum of the two— in this case it would be 12. However, if the same record matched the top category and its subcategory, then the count would be 11.

When no categories exist, the table still contains two rows. The top row, called All Records, is the total number of records. A second row, called Uncategorized, shows the number of documents/records that have yet to be categorized.

For each category in the pane, a small yellow bucket icon precedes the category name. If you double-click a category, or right-click in the tree table and select **Category Definitions**, the **Category Definitions** dialog box opens and presents all of the elements, called **descriptors**, that make up its definition, such as concepts, types, patterns, and category rules. For more information, see the topic “About Categories” on p. 103. By default, the category tree table does not show the descriptors in the categories. If you want to see the descriptors directly in the tree rather than in the **Category Definitions** dialog box, click the toggle button with the pencil icon in the toolbar. When this toggle button is selected, you can expand your tree to see the descriptors as well.

Scoring Categories

The **Docs.** column in the category tree table displays the number of records that are categorized into that specific category. If the numbers are out of date or are not calculated, an icon appears in that column. Keep in mind that the scoring process can take some time when you are working with larger datasets.

Selecting Categories in the Tree

When making selections in the tree, you can only select sibling categories — that is to say, if you select top level categories, you can not also select a subcategory. Or if you select 2 subcategories of a given category, you cannot simultaneously select a subcategory of another category. Selecting a discontinuous category will result in the loss of the previous selection.

Displaying in Data and Visualization Panes

When you select a row in the table, the **Visualization** and **Data** panes are refreshed automatically with information corresponding to your selection.

Refining Your Categories

Categorization may not yield perfect results for your data on the first try, and there may well be categories that you want to delete or combine with other categories. You may also find, through a review of the extraction results, that there are some categories that were not created that you would find useful. If so, you can make manual changes to the results to fine-tune them for your particular context. For more information, see the topic “Editing and Refining Categories” on p. 148.

- Edit or add to category definitions as well as move, merge, or delete categories. For more information, see “Editing and Refining Categories” below.
- Force specific response IDs into or out of categories. For more information, see the topic “Forcing Responses into Categories” on p. 153.
- Add text matches to categories to capture responses that contain the same text into the category. For more information, see the topic “Text Matching in Categories” on p. 154.
- Add category rules to a category to automatically classify responses into a category based on a logical expression. For more information, see the topic “Using Category Rules” on p. 138.
- Visualize how your categories work together. For more information, see the topic “Visualizing Graphs” in Chapter 7 on p. 159.
- Export your categorization results. For more information, see the topic “Exporting Categorization Results” in Chapter 4 on p. 52.

The Data Pane

As you create categories, there may be times when you might want to review some of the text data you are working with. For example, if you create a category in which 640 records are categorized, you might want to look at some or all of those records to see what text was actually written. You can review records in the Data pane, which is located in the lower right. If not visible by default, choose View > Panes > Data from the menus.

This pane presents, in table format, the response records for your open-ended data. Depending on what is selected in the other panes in this view, only the corresponding records appear in the pane. For example, if you select a concept in the Extraction Results pane, then only those records containing that concept (and associated terms) appear in the Data pane.

Whenever you select a concept or category in another pane and display the data, concepts (and associated terms) found in those records are highlighted in color to help you easily identify them in the text. The color coding corresponds to the types to which the concepts belong. You can also hover your mouse over color-coded items to display the concept under which it was extracted and the type to which it was assigned. Any text that was not extracted appears in black. Typically, these unextracted words are often connectors (*and* or *with*), pronouns (*me* or *they*), and verbs (*is*, *have*, or *take*).

Figure 6-3
Data pane

	Id	Response	Categories
1	14	I like that Product A has a lot of storage. Also, the interface is very easy to use.	memory device/memory
2	45	Everything! Product A rules! I can't wait to get a video one!	memory device/recording/video
3	68	Large storage capacity	...vice/memory/storage capacity
4	75	Small size. It has 512Mb of add-on memory, so it is quick to load and play music. It can also encode directly from external devices from the radio or a CD player.	...sumer electronics/home audio memory device/memory music radio size
5	76	storage capacity	...vice/memory/storage capacity
6	89	Small but lots of space (60 GB). Video is a bit of a toy but cool.	memory device/recording/video
7	98	Big storage capacity, also does video	...vice/memory/storage capacity memory device/recording/video
8	114	Large storage capacity and a good LCD screen for viewing digital photos	...vice/memory/storage capacity photos screen
9	120	The sound quality and the ability to record and mix my own play lists.	...nics/audio/sound/sound quality memory device/recording playlists
10	121	Storage capacity (40GB)	...vice/memory/storage capacity
11	146	It has a lot of storage capacity. I can fit a lot of songs on it. Also, it's very lightweight.	...vice/memory/storage capacity songs
	151	This has 256MB of memory; it holds about 50 songs. I've got another chip in my bag with another 50 songs on it. The cool thing about this is, it cost \$200 plus \$60 for a portable	music computer network listening

Note: To display all of the records for a given question in the Data pane, click the All Records node at the top of the Categories pane.

By default, the Data pane shows three columns (ID, Response, and Categories). However, you can add additional columns to this pane. The possible columns are as follows:

- **ID.** Lists the record or document identifier (ID) if one was imported.
- **Response.** Lists the text data from which concepts and type were extracted.
- **Categories.** Lists each of the categories to which a record belongs. Whenever this column is shown, refreshing the Data pane may take a bit longer so as to show the most up-to-date information. Categories are listed in this column according to their relevance to the record. For more information, see the topic “Category Relevance” on p. 97.
- **Force In.** Lists the categories into which you have forced a response. Responses can be forced into the category through the Edit > Force In menu selection. For more information, see the topic “Forcing Responses into Categories” on p. 153.
- **Force Out.** Lists the categories from which you have removed a response. Responses can be forced out of a category through the Edit > Force Out menu selection. Typically, this is used when a respondent’s sarcasm causes a response to be miscategorized. For more information, see the topic “Forcing Responses into Categories” on p. 153.
- **Text Match.** Lists any Text Matches found for each response. Strings can be defined to force specific text to be part of a category definition regardless of whether or not that string was extracted. For more information, see the topic “Text Matching in Categories” on p. 154.
- **Category Counts.** Provides the total number of categories to which the response belongs for this question.
- **Relevance Rank.** Provides a rank for each record in a single category. This rank shows how well the record fits into the category compared to the other records in that category. Select a category in the Categories pane (upper left pane) to see the rank in this column. For more information, see the topic “Category Relevance” on p. 97.
- **Response Flags.** Adds a column that shows any response flags you may be using. By clicking inside this column, you can change the type of flag that you assign to each response.
- *<any reference variable names>*. Adds a column for the selected reference variable to the Data pane. If you did not import any reference variables, none will be proposed here. A separate column is available for each reference variable. For more information, see the topic “Selecting Variables” in Chapter 3 on p. 32.

To Display Other Data Pane Columns

- ▶ From within the Data pane, right-click a column heading to open a context menu.
- ▶ From the menu choose Display Columns, and then select the column that you want to display in the Data pane. The new column appears in the pane.

Note: Forcing responses into and out of categories allows you to override the category definitions created by the automatic category building techniques without changing the actual category definition. For more information, see the topic “Forcing Responses into Categories” on p. 153.

Category Relevance

To help you build better categories, you can review the relevance of the records in each category as well as the relevance of all categories to which a record belongs.

Relevance of a Category to a Record

Whenever a record appears in the Data pane, all categories to which it belongs are listed in the Categories column. When a record belongs to multiple categories, the categories in this column appear in order from the most to the least relevant match. The category listed first is thought to correspond best to this record. For more information, see the topic “The Data Pane” on p. 95.

Relevance of a Record to a Category

When you select a category, you can review the relevance of each of its records in the Relevance Rank column in the Data pane. This relevance rank indicates how well the record fits into the selected category compared to the other records in that category. To see the rank of the records for a single category, select this category in the Categories pane (upper left pane) and the rank for record appears in the column. This column is not visible by default but you can choose to display it. For more information, see the topic “The Data Pane” on p. 95.

The lower the number for the record’s rank, the better the fit or the more relevant this record is to the selected category such that 1 is the best fit. If more than one record has the same relevance, each appears with the same rank followed by an equal sign (=) to denote they have equal relevance. For example, you might have the following ranks 1=, 1=, 3, 4, and so on, which means that there are two records that are equally considered as best matches for this category.

Tip: You could add the text of the most relevant record to the category annotation to help provide a better description of the category. Add the text directly from the Data pane by selecting the text and choosing Categories > Add to Annotation from the menus.

Figure 6-4
Data pane showing Categories and Relevance Rank

	Id	Response	Categories	Relevance Rank
1	6	Battery life. Portability. Accessories. Style.	Pos: Features-Design Pos: Quality-Reliability Pos: Usability	26=
2	7	I like its ability to store all of my music. I also like the ability to create playlists.	Pos: Features-Design Pos: General Satisfaction	141=
3	8	portability, capacity, sound quality, durability	Pos: Features-Design Pos: Quality-Reliability Pos: Storage Pos: Usability	89=
4	9	Small, great sound, capacity.	Pos: Features-Design Pos: Size-Weight Pos: Storage	23=
5	15	It holds a ton of music.	Pos: Features-Design	47=
6	17	its cool	Pos: Features-Design	47=
7	19	Others think it is cool and it sounds great.	Pos: Features-Design	14=
8	26	its great. i can share music with my friends and download tons of tunes off the internet.	Pos: Features-Design Pos: General Satisfaction	155=
9	29	Always having a good selection of music at hand	Pos: Features-Design	47=
10	31	Its portability enables me to listen to my music while I am milking cows and working in the fields.	Pos: Features-Design Pos: Usability	146=
11	32	it allows me to listen all of my music.	Pos: Features-Design	14=
12	34	The ability to build a playlist is the best feature. There are times I like to mix and match my selections.	Pos: Features-Design	146=
13	36	It has great sound quality. It also has capacity for all my music.	Pos: Features-Design Pos: Storage	23=
14	37	holds lots of music	Pos: Features-Design	47=
15	45	Everything! Product A rules! I can't wait to get a new one!	Pos: Features-Design	119=
16	46	Ability to carry large amounts of music in a small, lightweight device	Pos: Features-Design Pos: Size-Weight	26=

Methods and Strategies for Creating Categories

If you have not yet extracted or your extraction results are out of date, the use of one of the category building or extending techniques will prompt you for an extraction automatically. After you have applied a technique, the concepts and types that were grouped into a category are still available for category building with other techniques. This means that you may see a concept in multiple categories unless you choose not to reuse them.

In order to help you create the best categories, please review the following:

- **Methods for creating categories**
- **Strategies for creating categories**
- **Tips for creating categories**

Methods for Creating Categories

Because every dataset is unique, the number of category creation methods and the order in which you apply them may change over time. Additionally, since your text mining goals may be different from one set of data to the next, you may need to experiment with the different methods to see which one produces the best results for the given text data. None of the automatic techniques

will perfectly categorize your data; therefore we recommend finding and applying one or more automatic techniques that work well with your data.

Besides using text analysis packages (TAPs, *.tap) with prebuilt category sets, you can also categorize your responses using any combination of the following methods:

- **Automatic building techniques.** Several linguistic-based and frequency-based category options are available to automatically build categories for you. For more information, see the topic “Building Categories” on p. 105.
- **Automatic extending techniques.** Several linguistic techniques are available to extend existing categories by adding and enhancing descriptors so that they capture more records. For more information, see the topic “Extending Categories” on p. 120.
- **Manual techniques.** There are several manual methods, such as drag-and-drop. For more information, see the topic “Creating Categories Manually” on p. 124.
- **Code frames.** Import your own code frames, or copy/paste codes into the code frame manager. For more information, see the topic “Importing Predefined Categories” on p. 127.

Strategies for Creating Categories

The following list of strategies is by no means exhaustive but it can provide you with some ideas on how to approach the building of your categories.

- When you start a project, select a category set from a text analysis package (TAP) so that you begin your analysis with some prebuilt categories. These categories may sufficiently categorize your text right from the start. However, if you want to add more categories, you can edit the Build Categories settings (Categories > Build Settings). Open the Advanced Settings: Linguistics dialog and choose the Category input option Unused extraction results and build the additional categories.
- When you start a project, select a category set from a TAP. Next, drag and drop unused concepts or patterns into the categories as you deem appropriate. Then, extend the existing categories you’ve just edited (Categories > Extend Categories) to obtain more descriptors that are related to the existing category descriptors.
- Build categories automatically using the advanced linguistic settings (Categories > Build Categories). Then, refine the categories manually by deleting descriptors, deleting categories, or merging similar categories until you are satisfied with the resulting categories. Additionally, if you originally built categories **without** using the Generalize with wildcards where possible option, you can also try to simplify the categories automatically using the Extend Categories using the Generalize option.
- Import a predefined category file with very descriptive category names and/or annotations. Additionally, if you originally imported **without** choosing the option to import or generate descriptors from category names, you can later use the Extend Categories dialog and choose the Extend empty categories with descriptors generated from the category name. option. Then, extend those categories a second time but use the grouping techniques this time.
- Manually create a first set of categories by sorting concepts or concept patterns by frequency and then dragging and dropping the most interesting ones to the Categories pane. Once you have that initial set of categories, use the Extend feature (Categories > Extend Categories) to expand and refine all of the selected categories so they’ll include other related descriptors and thereby match more records.

After applying these techniques, we recommend that you review the resulting categories and use manual techniques to make minor adjustments, remove any misclassifications, or add records or words that may have been missed. Additionally, since using different techniques may produce redundant categories, you could also merge or delete categories as needed. For more information, see the topic “Editing and Refining Categories” on p. 148.

Tips for Creating Categories

In order to help you create better categories, you can review some tips that can help you make decisions on your approach.

Tips on Category-to-Response Ratio

When codes are created for a closed-end question, such as “*When was the last time you visited our outlet store?*”, the categories into which the responses fall should be mutually exclusive and exhaustive. That principle does not necessarily apply to qualitative text analysis, for at least two reasons:

- First, a general rule of thumb says that the longer the text record, the more distinct the ideas and opinions expressed. Thus, the chances that a record can be assigned multiple categories is greatly increased.
- Second, often there are various ways to group and interpret text records that are not logically separate. In the case of a survey with an open-ended question about the respondent’s political beliefs, we could create categories, such as *Liberal* and *Conservative*, or *Republican* and *Democrat*, as well as more specific categories, such as *Socially Liberal*, *Fiscally Conservative*, and so forth. These categories do not have to be mutually exclusive and exhaustive.

Tips on Number of Categories to Create

Except in the case of an extremely simple open-ended question, it is never intuitively obvious how many categories to create. The number of categories is not really an issue of concern. Instead, category creation should flow directly from the data—as you see something interesting with respect to the objectives of this survey, you can create a category to represent those attitudes and ideas.

- **Category frequency.** For a category to be useful, it has to contain a minimum number of records. One or two records may include something quite intriguing, but if they are one or two out of 1,000 records, the information they contain may not be frequent enough in the population to be practically useful.
- **Complexity.** The more categories you create, the more information you have to review and summarize after completing the analysis. However, too many categories, while adding complexity, may not add useful detail.

Unfortunately, there are no rules for determining how many categories are too many or for determining the minimum number of records per category. You will have to make such determinations based on the demands of your particular situation.

We can, however, offer advice about where to start. Although the number of categories should not be excessive, in the early stages of the analysis it is better to have too many rather than too few categories. It is easier to group categories that are relatively similar than to split off cases into new

categories, so a strategy of working from more to fewer categories is usually the best practice. Given the iterative nature of text mining and the ease with which it can be accomplished with this software program, building more categories is acceptable at the start.

Choosing the Best Descriptors

The following information contains some guidelines for choosing or making the best descriptors (concepts, types, TLA patterns, and category rules) for your categories. Descriptors are the building blocks of categories. When some or all of the text in a record matches a descriptor, the record is matched to the category.

Unless a descriptor contains or corresponds to an extracted concept or pattern, it will not be matched to any records. Therefore, use concepts, types, patterns, and category rules as described in the following paragraphs.

Since concepts represent not only themselves but also a set of underlying terms that can range from plural/singular forms, to synonyms, to spelling variations, only the concept itself should be used as a descriptor or as part of a descriptor. To learn more about the underlying terms for any given concept, click on the concept name in the Extraction Results pane. When you hover over the concept name, a tooltip appears and displays any of the underlying terms found in your text during the last extraction. Not all concepts have underlying terms. For example, if `car` and `vehicle` were synonyms but `car` was extracted as the concept with `vehicle` as an underlying term, then you only want to use `car` in a descriptor since it will automatically match records with `vehicle`.

Concepts and Types as Descriptors

Use a concept as a descriptor when you want to find all records containing that concept (or any of its underlying terms). In this case, the use of a more complex category rule is not needed since the exact concept name is sufficient. Keep in mind that when you use resources that extract opinions, sometimes concepts can change during TLA pattern extraction to capture the truer sense of the sentence (refer to the example in the next section on TLA).

For example, a survey response indicating each person's favorite fruits such as "*Apple and pineapple are the best*" could result in the extraction of `apple` and `pineapple`. By adding the concept `apple` as a descriptor to your category, all responses containing the concept `apple` (or any of its underlying terms) are matched to that category.

However, if you are interested in simply knowing which responses mention *apple* in any way, you can write a category rule such as `* apple *` and you will also capture responses that contain concepts such as `apple`, `apple sauce`, or `french apple tart`.

You can also capture all the records that contain concepts that were typed the same way by using a type as a descriptor directly such as `<Fruit>`. Please note that you cannot use `*` with types.

For more information, see the topic "Extracted Results: Concepts, Types, and Patterns" in Chapter 5 on p. 77.

Text Link Analysis (TLA) Patterns as Descriptors

Use a TLA pattern result as a descriptor when you want to capture finer, nuanced ideas. When text is analyzed during TLA extraction, the text is processed one sentence, or clause, at a time rather than looking at the entire text (the record). By considering all of the parts of a single

sentence together, TLA can identify opinions, relationships between two elements, or a negation, for example, and understand the truer sense. You can use concept patterns or type patterns as descriptors.

For example, if we had the text “*the room was not that clean*”, the following concepts could be extracted: `room` and `clean`. However, if TLA extraction was enabled in the extraction setting, TLA could detect that `clean` was used in a negative way and actually corresponds to `not clean`, which is a synonym of the concept `dirty`. Here, you can see that using the concept `clean` as a descriptor on its own would match this text but could also capture other or records mentioning cleanliness. Therefore, it might be better to use the TLA concept pattern with `dirty` as output concept since it would match this text and likely be a more appropriate descriptor.

Category Business Rules as Descriptors

Category rules are statements that automatically classify records into a category based on a logical expression using extracted concepts, types, and patterns as well as Boolean operators. For example, you could write an expression that means *include all records that contain the extracted concept embassy but not argentina in this category*.

You can write and use category rules as descriptors in your categories to express several different ideas using `&`, `|`, and `!()` Booleans. For detailed information on the syntax of these rules and how to write and edit them, see “Using Category Rules” on p. 138

- Use a category rule with the `&` (AND) Boolean operator to help you find records in which 2 or more concepts occur. The 2 or more concepts connected by `&` operators do not need to occur in the same sentence or phrase, but can occur anywhere in the same record to be considered a match to the category. For example, if you create the category rule `food & cheap` as a descriptor, it would match a record containing the text, “*the food was pretty expensive, but the rooms were cheap*” despite the fact that `food` was not the noun being called `cheap` since the text contained both `food` and `cheap`.
- Use a category rule with the `!()` (NOT) Boolean operator as a descriptor to help you find records in which some things occur but others do not. This can help avoid grouping information that may seem related based on words but not on context. For example, if you create the category rule `<Organization> & !(ibm)` as a descriptor, it would match the following text *SPSS Inc. was a company founded in 1967* and not match the following text *the software company was acquired by IBM.*
- Use a category rule with the `|` (OR) Boolean operator as a descriptor to help you find records containing one of several concepts or types. For example, if you create the category rule `(personnel|staff|team|coworkers) & bad` as a descriptor, it would match any records in which any of those nouns are found with the concept `bad`.
- Use types in category rules to make them more generic and possibly more deployable. For example, if you were working with hotel data, you might be very interested in learning what customers think about hotel personnel. Related terms might include words such as `receptionist`, `waiter`, `waitress`, `reception desk`, `front desk` and so on. You could, in this case, create a new type called `<HotelStaff>` and add all of the preceding terms to that type. While it is possible to create one category rule for every kind of staff such as `[* waitress * & nice]`, `[* desk * & friendly]`, `[* receptionist * & accommodating]`, you could create a single, more generic category rule using the `<HotelStaff>` type to capture

all responses that have favorable opinions of the hotel staff in the form of [`<HotelStaff>` & `<Positive>`].

Note: You can use both + and & in category rules when including TLA patterns in those rules. For more information, see the topic “Using TLA Patterns in Category Rules” on p. 140.

Example of how concepts, TLA, or category rules as descriptors match differently

The following example demonstrates how using a concept as a descriptor, category rule as a descriptor, or using a TLA pattern as a descriptor affects how records are categorized. Let’s say you had the following 5 records.

- A: “*awesome restaurant staff, excellent food and rooms comfortable and clean.*”
- B: “*restaurant personnel was awful, but rooms were clean.*”
- C: “*Comfortable, clean rooms.*”
- D: “*My room was not that clean.*”
- E: “*Clean.*”

Since the records include the word *clean* and you want to capture this information, you could create one of the descriptors shown in the following table. Based on the essence you are trying to capture, you can see how using one kind of descriptor over another can produce different results.

Table 6-1
How Example Records Matched Descriptors

Descriptor	A	B	C	D	E	Explanation
clean	match	match	match	match	match	Descriptor is an extracted concept. Every record contained the concept <code>clean</code> , even record D since without TLA, it is not known automatically that “ <i>not clean</i> ” means <i>dirty</i> by the TLA rules.
clean + .	-	-	-	-	match	Descriptor is a TLA pattern that represents <code>clean</code> by itself. Matched only the record where <code>clean</code> was extracted with no associated concept during TLA extraction.
[clean]	match	match	match	-	match	Descriptor is a category rule that looks for a TLA rule that contains <code>clean</code> on its own or with something else. Matched all records where a TLA output containing <code>clean</code> was found regardless of whether <code>clean</code> was linked to another concept such as <code>room</code> and in any slot position.

About Categories

Categories refer to a group of closely related concepts, opinions, or attitudes. To be useful, a category should also be easily described by a short phrase or label that captures its essential meaning.

For example, if you are analyzing survey responses from consumers about a new laundry soap, you can create a category labeled *odor* that contains all of the responses describing the smell of the product. However, such a category would not differentiate between those who found the smell pleasant and those who found it offensive. Since IBM® SPSS® Text Analytics for Surveys is capable of extracting opinions when using the appropriate resources, you could then create two other categories to identify respondents who *enjoyed the odor* and respondents who *disliked the odor*.






You can create and work with your categories in the Categories pane in the upper left pane of the text analysis window. Each category is defined by one or more descriptors. **Descriptors** are concepts, types, and patterns, as well as category rules that have been used to define a category.

If you want to see the descriptors that make up a given category, you can expand the category in the tree. Icons are shown in the tree so that you can easily identify each element. Only the first level defines the category. If you expand the definitions further, you can see examples of what was found in the data.

When you build categories automatically using category building techniques such as concept inclusion, the techniques will use concepts and types as the descriptors to create your categories. You can also add patterns or parts of those patterns as category descriptors. Lastly, you can manually create category rules to use as descriptors in your categories. For more information, see the topic “Using Category Rules” on p. 138.

For example, if you add a type to a category definition, any concepts assigned to that type would automatically be included, even if only a handful are present in the data at this time. This helps when reusing category definitions with new data. For more information, see the topic “Copying Categories” on p. 156. You can also manually create category rules to use as descriptors in your categories. For more information, see the topic “Using Category Rules” on p. 138.

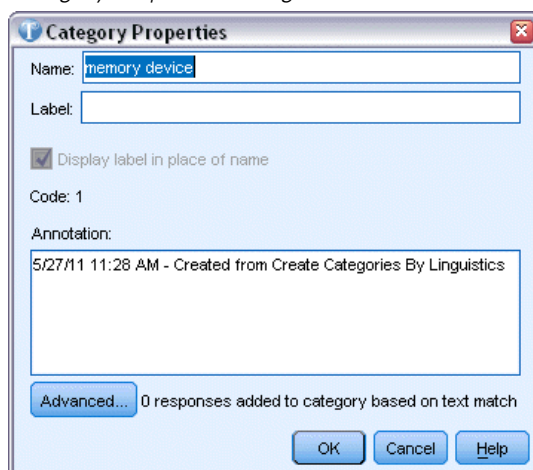
Table 6-2
Icons to identify elements in definitions

Icon	Description
	Concept.
	Type, which can be expanded to see the concepts it contains.
	Concept pattern, which can be expanded to see the specific concepts in patterns.
	Type pattern, which can be expanded to the concept pattern level.
	Category rules in the category. Right-click the rule name to edit the rule.

Category Properties

In addition to descriptors, categories also have properties you can edit in order to rename categories, add a label, or add an annotation, or access the text matching dialog.

Figure 6-5
Category Properties dialog



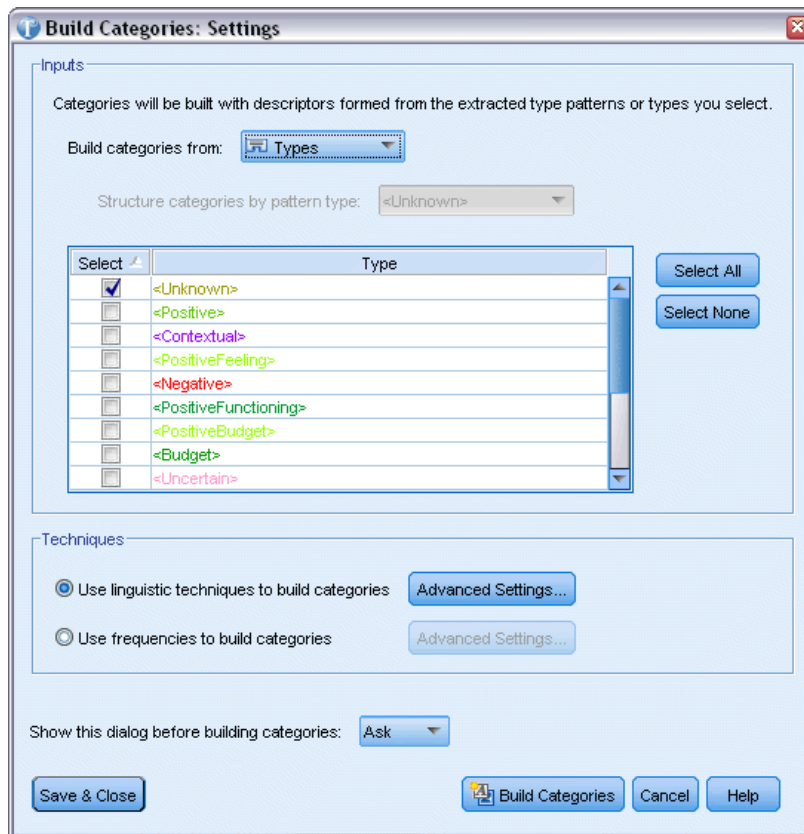
The following properties exist:

- **Name.** This name appears in the tree by default. When a category is created using an automated technique, it is given a name automatically.
- **Label.** Using labels is helpful in creating more meaningful category descriptions for use in other products or in other tables or graphs. If you choose the option to display the label, then the label is used in the interface to identify the category. In IBM® SPSS® Statistics, they are exported as variable labels. In Microsoft Excel, they are exported as a separate row.
- **Code.** The code number corresponds to the code value for this category. You edit this code in the Code Frame Manager. The Code Frame Manager allows you to edit the name, label, and code for each category as well as copy and paste entire code frames. .
- **Annotation.** You can add a short description for each category in this field. When a category is generated by the Build Categories dialog, a note is added to this annotation automatically. You can also add sample text to an annotation directly from the Data pane by selecting the text and choosing Categories > Add to Annotation from the menus.
- **Advanced text match.** Using the Advanced button, you can add words or phrases to a category definition. Often this is used to override a missed extraction. For more information, see the topic “Text Matching in Categories” on p. 154.

Building Categories

While you may have categories from a text analysis package, you can also build categories automatically using a number of linguistic and frequency techniques. Through the Build Categories Settings dialog box, you can apply the automated linguistic and frequency techniques to produce categories from either concepts or from concept patterns.

Figure 6-6
Build Categories dialog box



In general, categories can be made up of different kinds of descriptors (types, concepts, TLA patterns, category rules). When you build categories using the automated category building techniques, the resulting categories are named after a concept or concept pattern (depending on the input you select) and each contains a set of descriptors. These descriptors may be in the form of category rules or concepts and include all the related concepts discovered by the techniques.

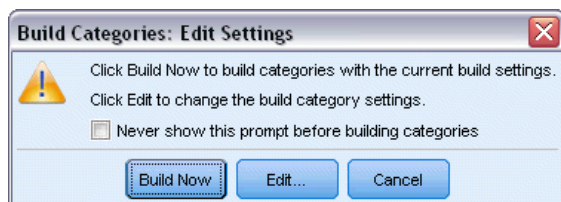
After building categories, you can learn a lot about the categories by reviewing them in the Categories pane or exploring them through the graphs and charts. You can then use manual techniques to make minor adjustments, remove any misclassifications, or add records or words that may have been missed. After you have applied a technique, the concepts, types, and patterns that were grouped into a category are still available for other techniques. Also, since using different techniques may also produce redundant or inappropriate categories, you can also merge or delete categories. For more information, see the topic “Editing and Refining Categories” on p. 148.

Important! In earlier releases, co-occurrence and synonym rules were surrounded by square brackets. In this release, square brackets now indicate a pattern result. Instead, co-occurrence and synonym rules will be encapsulated by parentheses such as (speaker systems | speakers).

To Build Categories

- From the menus, choose Categories > Build Categories. Unless you have chosen to never prompt, a message box appears.

Figure 6-7
Prompting before building



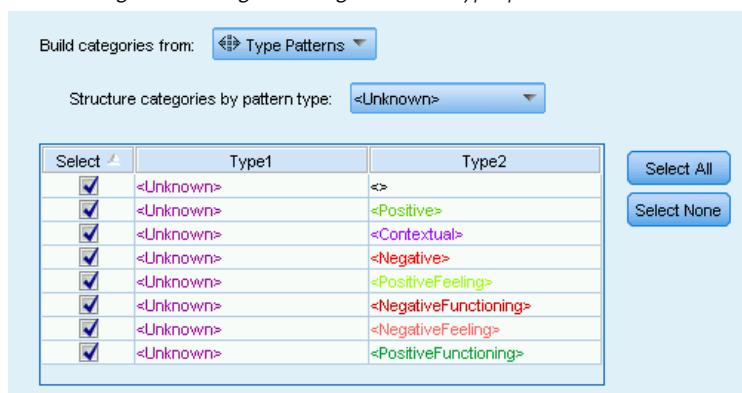
- ▶ Choose whether you want to build now or edit the settings first.
 - Click Build Now to begin building categories using the current settings. The settings selected by default are often sufficient to begin the categorization process. The category building process begins and a progress dialog appears.
 - Click Edit to review and modify the build settings.

Inputs

The categories are built from descriptors derived from either type patterns or types. By default, type patterns are selected in the dropdown list. In the table, you can select the individual types or patterns to include in the category building process.

Type patterns. If you select type patterns, categories are built from patterns rather than types and concepts on their own. In that way, any records containing a concept pattern belonging to the selected type pattern are categorized. So, if you select the <Budget> and <Positive> type pattern in the table, categories such as *cost & <Positive>* or *rates & excellent* could be produced. The table displays only one row for each type combination such as <Location> + <Positive> and <Positive> + <Location>, and their order is unimportant to how the categories are generated.

Figure 6-8
Build categories dialog showing available type patterns



When using type patterns as input for automated category building, there are times when the techniques identify multiple ways to form the category structure. Technically, there is no single right way to produce the categories; however you might find one structure more suited to your analysis than another. To help customize the output in this case, you can designate a type as the preferred focus. Choose this type in the Structure categories by pattern type: field and the table will

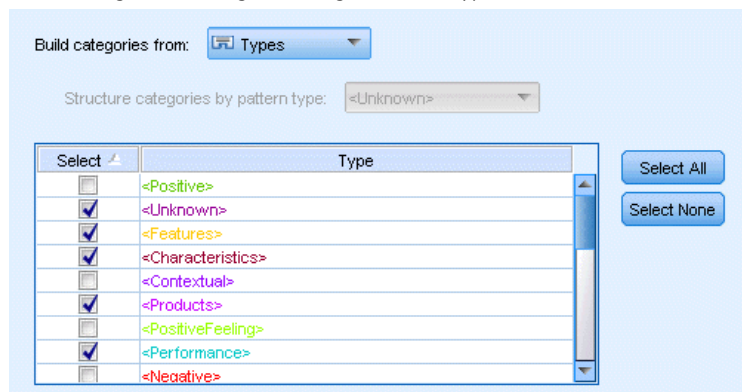
be updated to show only the applicable patterns containing the selected type. More often than not, <Unknown> will be preselected for you. This results in all of the patterns containing the type <Unknown> being selected. The table displays the types in descending order starting with the one with the greatest number of records.

Types. If you select types, the categories will be built from the concepts belonging to the selected types. So if you select the <Budget> type in the table, categories such as *cost* or *price* could be produced since *cost* and *price* are concepts assigned to the <Budget> type.

By default, only the types that capture the most records are selected. This pre-selection allows you to quickly focus in on the most interesting types and avoid building uninteresting categories. The table displays the types in descending order starting with the one with the greatest number of records. Types from the Opinions library are deselected by default in the types table.

The input you choose affects the categories you obtain. When you choose to use Types as input, you can see the clearly related concepts more easily. For example, if you build categories using Types as input, you could obtain a category *Fruit* with concepts such as *apple*, *pear*, *citrus fruits*, *orange* and so on. If you choose Type Patterns as input instead and select the pattern <Unknown> + <Positive>, for example, then you might get a category *fruit* + <Positive> with one or two kinds of fruit such as *fruit* + *tasty* and *apple* + *good*. This second result only shows 2 concept patterns because the other occurrences of fruit are not necessarily positively qualified. And while this might be good enough for your current text data, in longitudinal studies where you use different document sets, you may want to manually add in other descriptors such as *citrus fruit* + *positive* or use types. Using types alone as input will help you to find all possible fruit.

Figure 6-9
Build categories dialog showing available types



Techniques

Because every dataset is unique, the number of methods and the order in which you apply them may change over time. Since your text mining goals may be different from one set of data to the next, you may need to experiment with the different techniques to see which one produces the best results for the given text data.

You do not need to be an expert in these settings to use them. By default, the most common and average settings are already selected. Therefore, you can bypass the advanced setting dialogs and go straight to building your categories. Likewise, if you make changes here, you do not have to come back to the settings dialog each time since the latest settings are always retained.

Select either the linguistic or frequency techniques and click the Advanced Settings button to display the settings for the techniques selected. None of the automatic techniques will perfectly categorize your data; therefore we recommend finding and applying one or more automatic techniques that work well with your data. You cannot build using linguistic and frequency techniques simultaneously.

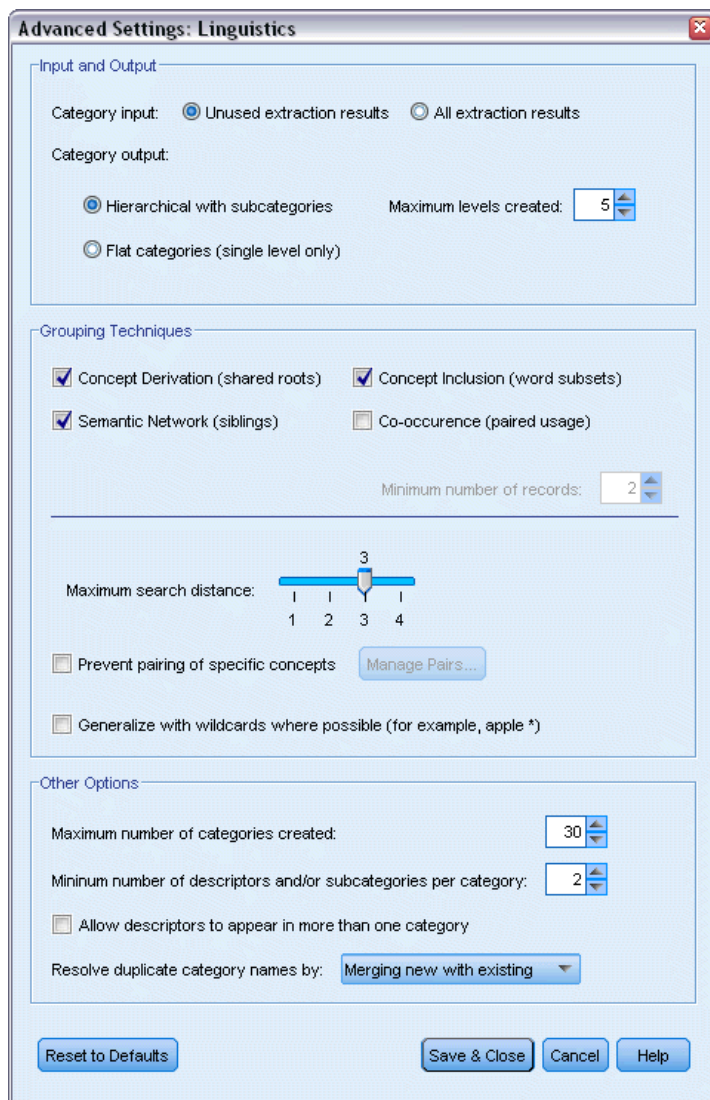
- **Advanced linguistic techniques.** For more information, see on p. 109.
- **Advanced frequency techniques.** For more information, see on p. 118.

Advanced Linguistic Settings

When you build categories, you can select from a number of advanced linguistic category building techniques including *concept root derivation*, *concept inclusion*, *semantic networks* (English text only), and *co-occurrence rules*. These techniques can be used individually or in combination with each other to create categories.

Keep in mind that because every dataset is unique, the number of methods and the order in which you apply them may change over time. Since your text mining goals may be different from one set of data to the next, you may need to experiment with the different techniques to see which one produces the best results for the given text data. None of the automatic techniques will perfectly categorize your data; therefore we recommend finding and applying one or more automatic techniques that work well with your data.

Figure 6-10
Advanced Settings: Linguistics dialog box for building categories



Input and Output

Category input. Select from what the categories will be built:

- **Unused extraction results.** This option enables categories to be built from extraction results that are not used in any existing categories. This minimizes the tendency for records to match multiple categories and limits the number of categories produced.
- **All extraction results.** This option enables categories to be built using any of the extraction results. This is most useful when no or few categories already exist.

Category output. Select the general structure for the categories that will be built:

- Hierarchical with subcategories. This option enables the creation of subcategories and sub-subcategories. You can set the depth of your categories by choosing the maximum number of levels (Maximum levels created field) that can be created. If you choose 3, categories could contain subcategories and those subcategories could also have subcategories.
- Flat categories (single level only). This option enables only one level of categories to be built, meaning that no subcategories will be generated.

Grouping Techniques

Each of the techniques available is well suited to certain types of data and situations, but often it is helpful to combine techniques in the same analysis to capture the full range of records. You may see a concept in multiple categories or find redundant categories.

Concept Root Derivation. This technique creates categories by taking a concept and finding other concepts that are related to it by analyzing whether any of the concept components are morphologically related, or share roots. This technique is very useful for identifying synonymous compound word concepts, since the concepts in each category generated are synonyms or closely related in meaning. It works with data of varying lengths and generates a smaller number of compact categories. For example, the concept `opportunities to advance` would be grouped with the concepts `opportunity for advancement` and `advancement opportunity`. For more information, see the topic “Concept Root Derivation” on p. 114.

Semantic Network. This technique begins by identifying the possible senses of each concept from its extensive index of word relationships and then creates categories by grouping related concepts. This technique is best when the concepts are known to the semantic network and are not too ambiguous. It is less helpful when text contains specialized terminology or jargon unknown to the network. In one example, the concept `granny smith apple` could be grouped with `gala apple` and `winesap apple` since they are siblings of the `granny smith`. In another example, the concept `animal` might be grouped with `cat` and `kangaroo` since they are hyponyms of `animal`. This technique is available for English text only in this release. For more information, see the topic “Semantic Networks” on p. 116.

Concept Inclusion. This technique builds categories by grouping multiterm concepts (compound words) based on whether they contain words that are subsets or supersets of a word in the other. For example, the concept `seat` would be grouped with `safety seat`, `seat belt`, and `seat belt buckle`. For more information, see the topic “Concept Inclusion” on p. 115.

Co-occurrence. This technique creates categories from co-occurrences found in the text. The idea is that when concepts or concept patterns are often found together in documents and records, that co-occurrence reflects an underlying relationship that is probably of value in your category definitions. When words co-occur significantly, a co-occurrence rule is created and can be used as a category descriptor for a new subcategory. For example, if many records contain the words `price` and `availability` (but few records contain one without the other), then these concepts could be

grouped into a co-occurrence rule, (`price & available`) and assigned to a subcategory of the category `price` for instance. For more information, see the topic “Co-occurrence Rules” on p. 117.

- **Minimum number of records.** To help determine how interesting co-occurrences are, define the minimum number of records that must contain a given co-occurrence for it to be used as a descriptor in a category.

Maximum search distance. Select how far you want the techniques to search before producing categories. The lower the value, the fewer results you will get—however, these results will be less noisy and are more likely to be significantly linked or associated with each other. The higher the value, the more results you might get—however, these results may be less reliable or relevant. While this option is globally applied to all techniques, its effect is greatest on co-occurrences and semantic networks.

Prevent pairing of specific concepts. Select this checkbox to stop the process from grouping or pairing two concepts together in the output. To create or manage concept pairs, click *Manage Pairs...* For more information, see the topic “Managing Link Exception Pairs” on p. 113.

Generalize with wildcards where possible. Select this option to allow the product to generate generic rules in categories using the asterisk wildcard. For example, instead of producing multiple descriptors such as [`apple tart + .`] and [`apple sauce + .`], using wildcards might produce [`apple * + .`]. If you generalize with wildcards, you will often get exactly the same number of records as you did before. However, this option has the advantage of reducing the number and simplifying category descriptors. Additionally, this option increases the ability to categorize more records using these categories on new text data (for example, in longitudinal/wave studies).

Other Options for Building Categories

In addition to selecting the grouping techniques to apply, you can edit several other build options as follow:

Maximum number of categories created. Use this option to limit the number of categories that can be generated when you click the *Build Categories* button next. In some cases, you might get better results if you set this value high and then delete any of the uninteresting categories.

Minimum number of descriptors and/or subcategories per category. Use this option to define the minimum number of descriptors and subcategories a category must contain in order to be created. This option helps limit the creation of categories that do not capture a significant number of records .

Allow descriptors to appear in more than one category. When selected, this option allows descriptors to be used in more than one of the categories that will be built next. This option is generally selected since items commonly or “naturally” fall into two or more categories and allowing them to do so usually leads to higher quality categories. If you do not select this option, you reduce the overlap of records in multiple categories and depending on the type of data you have, this might be desirable. However, with most types of data, restricting descriptors to a single category usually results in a loss of quality or category coverage. For example, let’s say you had the concept `car seat manufacturer`. With this option, this concept could appear in one category based on the text `car seat` and in another one based on `manufacturer`. But if this option is not selected, although you may still get both categories, the concept `car seat manufacturer` will only

appear as a descriptor in the category it best matches based on several factors including the number of records in which `car seat` and `manufacturer` each occur.

Resolve duplicate category names by. Select how to handle any new categories or subcategories whose names would be the same as existing categories. You can either merge the new ones (and their descriptors) with the existing categories with the same name. Alternatively, you can choose to skip the creation of any categories if a duplicate name is found in the existing categories.

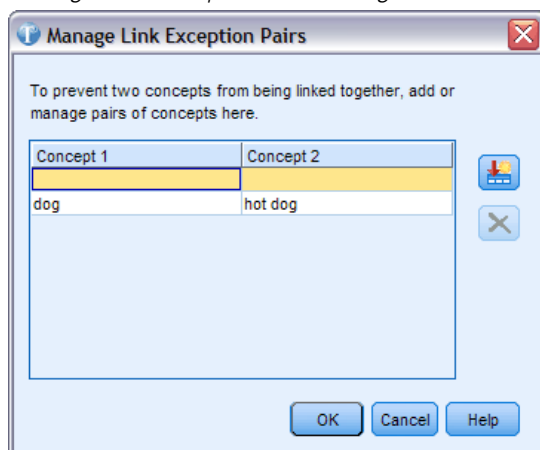
Managing Link Exception Pairs

During category building, and concept mapping, the internal algorithms group words by known associations. To prevent two concepts from being paired, or linked together, you can turn on this feature in Build Categories Advanced Settings dialog, and Concept Map Index Settings dialog and click the Manage Pairs button.

In the resulting Manage Link Exceptions dialog, you can add, edit, or delete concept pairs. Enter one pair per line. Entering pairs here will prevent the pairing from occurring when building or extending categories, and concept mapping. Enter words exactly as you want them, for example the accented version of word is not equal to the unaccented version of the word.

For example, if you wanted to make sure that `hot dog` and `dog` are not grouped, you could add the pair as a separate line in the table.

Figure 6-11
Manage Link Exception Pairs dialog



About Linguistic Techniques

When you build or extend you categories, you can select from a number of advanced linguistic category building techniques including *concept root derivation*, *concept inclusion*, *semantic networks* (English only), and *co-occurrence rules*. These techniques can be used individually or in combination with each other to create categories.

You do not need to be an expert in these settings to use them. By default, the most common and average settings are already selected. If you want, you can bypass this advanced setting dialog and go straight to building or extending your categories. Likewise, if you make changes here, you do not have to come back to the settings dialog each time since it will remember what you last used.

However, keep in mind that because every dataset is unique, the number of methods and the order in which you apply them may change over time. Since your text mining goals may be different from one set of data to the next, you may need to experiment with the different techniques to see which one produces the best results for the given text data. None of the automatic techniques will perfectly categorize your data; therefore we recommend finding and applying one or more automatic techniques that work well with your data.

The main automated linguistic techniques for category building are:

- **Concept root derivation.** This technique creates categories by taking a concept and finding other concepts that are related to it through analyzing whether any of the concept components are morphologically related. For more information, see the topic “Concept Root Derivation” on p. 114.
- **Concept inclusion.** This technique creates categories by taking a concept and finding other concepts that include it. For more information, see the topic “Concept Inclusion” on p. 115.
- **Semantic network.** This technique begins by identifying the possible senses of each concept from its extensive index of word relationships and then creates categories by grouping related concepts. For more information, see the topic “Semantic Networks” on p. 116. This option is only available for English text.
- **Co-occurrence.** This technique creates co-occurrence rules that can be used to create a new category, extend a category, or as input to another category technique. For more information, see the topic “Co-occurrence Rules” on p. 117.

Concept Root Derivation

The concept root derivation technique creates categories by taking a concept and finding other concepts that are related to it through analyzing whether any of the concept components are morphologically related. A component is a word. The technique attempts to group concepts by looking at the endings (suffixes) of each component in a concept and finding other concepts that could be derived from them. The idea is that when words are derived from each other, they are likely to share or be close in meaning. In order to identify the endings, internal language-specific rules are used. For example, the concept `opportunities to advance` would be grouped with the concepts `opportunity for advancement` and `advancement opportunity`.

You can use concept root derivation on any sort of text. By itself, it produces fairly few categories, and each category tends to contain few concepts. The concepts in each category are either synonyms or situationally related. You may find it helpful to use this algorithm even if you are building categories manually; the synonyms it finds may be synonyms of those concepts you are particularly interested in.

Note: You can prevent concepts from being grouped together by specifying them explicitly. For more information, see the topic “Managing Link Exception Pairs” on p. 113.

Term Componentization and De-inflecting

When the concept root derivation or the concept inclusion techniques are applied, the terms are first broken down into components (words) and then the components are de-inflected. When a technique is applied, the concepts and their associated terms are loaded and split into components

based on separators, such as spaces, hyphens, and apostrophes. For example, the term `system administrator` is split into components such as `{administrator, system}`.

However, some parts of the original term may not be used and are referred to as stop words. In English, some of these ignorable components might include `a`, `and`, `as`, `by`, `for`, `from`, `in`, `of`, `on`, `or`, `the`, `to`, and `with`.

For example, the term `examination of the data` has the component set `{data, examination}`, and both `of` and `the` are considered ignorable. Additionally, component order is not in a component set. In this way, the following three terms could be equivalent: `cough relief for child`, `child relief from a cough`, and `relief of child cough` since they all have the same component set `{child, cough, relief}`. Each time a pair of terms are identified as being equivalent, the corresponding concepts are merged to form a new concept that references all of the terms.

Additionally, since the components of a term may be inflected, language-specific rules are applied internally to identify equivalent terms regardless of inflectional variation, such as plural forms. In this way, the terms `level of support` and `support levels` can be identified as equivalent since the de-inflected singular form would be `level`.

How Concept Root Derivation Works

After terms have been componentized and de-inflected (see previous section), the concept root derivation algorithm analyzes the component endings, or suffixes, to find the component root and then groups the concepts with other concepts that have the same or similar roots. The endings are identified using a set of linguistic derivation rules specific to the text language. For example, there is a derivation rule for English language text that states that a concept component ending with the suffix `ical` might be derived from a concept having the same root stem and ending with the suffix `ic`. Using this rule (and the de-inflection), the algorithm would be able to group the concepts `epidemiologic study` and `epidemiological studies`.

Since terms are already componentized and the ignorable components (for example, `in` and `of`) have been identified, the concept root derivation algorithm would also be able to group the concept `studies in epidemiology` with `epidemiological studies`.

The set of component derivation rules has been chosen so that most of the concepts grouped by this algorithm are synonyms: the concepts `epidemiologic studies`, `epidemiological studies`, `studies in epidemiology` are all equivalent terms. To increase completeness, there are some derivation rules that allow the algorithm to group concepts that are situationally related. For example, the algorithm can group concepts such as `empire builder` and `empire building`.

Concept Inclusion

The concept inclusion technique builds categories by taking a concept and, using lexical series algorithms, identifies concepts included in other concepts. The idea is that when words in a concept are a subset of another concept, it reflects an underlying semantic relationship. Inclusion is a powerful technique that can be used with any type of text.

This technique works well in combination with semantic networks but can be used separately. Concept inclusion may also give better results when the records contain lots of domain-specific terminology or jargon. This is especially true if you have tuned the dictionaries beforehand so that the special terms are extracted and grouped appropriately (with synonyms).

How Concept Inclusion Works

Before the concept inclusion algorithm is applied, the terms are componentized and de-inflected. For more information, see the topic “Concept Root Derivation” on p. 114. Next, the concept inclusion algorithm analyzes the component sets. For each component set, the algorithm looks for another component set that is a subset of the first component set.

For example, if you have the concept `continental breakfast`, which has the component set `{breakfast, continental}`, and you have the concept `breakfast`, which has the component set `{breakfast}`, the algorithm would conclude that `continental breakfast` is a kind of `breakfast` and group these together.

In a larger example, if you have the concept `seat` in the Extraction Results pane and you apply this algorithm, then concepts such as `safety seat`, `leather seat`, `seat belt`, `seat belt buckle`, `infant seat carrier`, and `car seat laws` would also be grouped in that category.

Since terms are already componentized and the ignorable components (for example, `in` and `of`) have been identified, the concept inclusion algorithm would recognize that the concept `advanced spanish course` includes the concept `course in spanish`.

Note: You can prevent concepts from being grouped together by specifying them explicitly. For more information, see the topic “Managing Link Exception Pairs” on p. 113.

Semantic Networks

In this release, the semantic networks technique is only available for English language text.

This technique builds categories using a built-in network of word relationships. For this reason, this technique can produce very good results when the terms are concrete and are not too ambiguous. However, you should not expect the technique to find many links between highly technical/specialized concepts. When dealing with such concepts, you may find the concept inclusion and concept root derivation techniques to be more useful.

How Semantic Network Works

The idea behind the semantic network technique is to leverage known word relationships to create categories of synonyms or hyponyms. A **hyponym** is when one concept is a sort of second concept such that there is a hierarchical relationship, also known as an ISA relationship. For example, if `animal` is a concept, then `cat` and `kangaroo` are hyponyms of `animal` since they are sorts of animals.

In addition to synonym and hyponym relationships, the semantic network technique also examines part and whole links between any concepts from the `<Location>` type. For example, the technique will group the concepts `normandy`, `provence`, and `france` into one category because Normandy and Provence are parts of France.

Semantic networks begin by identifying the possible senses of each concept in the semantic network. When concepts are identified as synonyms or hyponyms, they are grouped into a single category. For example, the technique would create a single category containing these three concepts: `eating apple`, `dessert apple`, and `granny smith` since the semantic network contains the information that: 1) `dessert apple` is a synonym of an `eating apple`, and 2) `granny smith` is a sort of `eating apple` (meaning it is a hyponym of `eating apple`).

Taken individually, many concepts, especially uniterms, are ambiguous. For example, the concept `buffet` can denote a sort of meal or a piece of furniture. If the set of concepts includes `meal`, `furniture` and `buffet`, then the algorithm is forced to choose between grouping `buffet` with `meal` or with `furniture`. Be aware that in some cases the choices made by the algorithm may not be appropriate in the context of a particular set of records.

The semantic network technique can outperform concept inclusion with certain types of data. While both the semantic network and concept inclusion recognize that `apple pie` is a sort of `pie`, only the semantic network recognizes that `tart` is also a sort of `pie`.

Semantic networks will work in conjunction with the other techniques. For example, suppose that you have selected both the semantic network and inclusion techniques and that the semantic network has grouped the concept `teacher` with the concept `tutor` (because a `tutor` is a kind of `teacher`). The inclusion algorithm can group the concept `graduate tutor` with `tutor` and, as a result, the two algorithms collaborate to produce an output category containing all three concepts: `tutor`, `graduate tutor`, and `teacher`.

Options for Semantic Network

There are a number of additional settings that might be of interest with this technique.

- Change the Maximum search distance. Select how far you want the techniques to search before producing categories. The lower the value, the fewer results produced—however, these results will be less noisy and are more likely to be significantly linked or associated with each other. The higher the value, the more results you will get—however, these results may be less reliable or relevant.

For example, depending on the distance, the algorithm searches from `Danish pastry` up to `coffee roll` (its parent), then `bun` (grand parent) and on upwards to `bread`.

By reducing the search distance, this technique produces smaller categories that might be easier to work with if you feel that the categories being produced are too large or group too many things together.

Important! Additionally, we recommend that you do not apply the option `Accommodate spelling errors` for a minimum root character limit of (defined in the `Extract` dialog box) for fuzzy grouping when using this technique since some false groupings can have a largely negative impact on the results.

Co-occurrence Rules

Co-occurrence rules enable you to discover and group concepts that are strongly related within the set of records. The idea is that when concepts are often found together in records, that co-occurrence reflects an underlying relationship that is probably of value in your category definitions. This technique creates co-occurrence rules that can be used to create a new category,

extend a category, or as input to another category technique. Two concepts strongly co-occur if they frequently appear together in a set of records and rarely separately in any of the other records. This technique can produce good results with larger datasets with at least several hundred records.

For example, if many records contain the words `price` and `availability`, these concepts could be grouped into a co-occurrence rule, `(price & available)`. In another example, if the concepts `peanut butter`, `jelly`, `sandwich` and appear more often together than apart, they would be grouped into a concept co-occurrence rule `(peanut butter&jelly & sandwich)`.

Important! In earlier releases, co-occurrence and synonym rules were surrounded by square brackets. In this release, square brackets now indicate a pattern result. Instead, co-occurrence and synonym rules will be encapsulated by parentheses such as `(speaker systems|speakers)`.

How Co-occurrence Rules Works

This technique scans the records looking for two or more concepts that tend to appear together. Two or more concepts strongly co-occur if they frequently appear together in a set of records and if they seldom appear separately in any of the other records.

When co-occurring concepts are found, a category rule is formed. These rules consist of two or more concepts connected using the `&` Boolean operator. These rules are logical statements that will automatically classify a record into a category if the set of concepts in the rule all co-occur in that record.

Options for Co-occurrence Rules

If you are using the co-occurrence rule technique, you can fine-tune several settings that influence the resulting rules:

- **Change the Maximum search distance.** Select how far you want the techniques to search before producing categories. The lower the value, the fewer results produced—however, these results will be less noisy and are more likely to be significantly linked or associated with each other. The higher the value, the more results you will get—however, these results may be less reliable or relevant. When working on co-occurrences, the default value for the search distance results in many, many co-occurrences of which many are weakly linked and hence uninteresting. When you reduce the search distance, you filter out the weaker co-occurrences and obtain the more significant results.
- **Minimum number of records.** To help determine how interesting co-occurrences are, define the minimum number of records that must contain a given co-occurrence for it to be used as a descriptor in a category. With smaller data sets, the lower you set this option, the easier it will be to find co-occurrences.

Note: You can prevent concepts from being grouped together by specifying them explicitly. For more information, see the topic “Managing Link Exception Pairs” on p. 113.

Advanced Frequency Settings

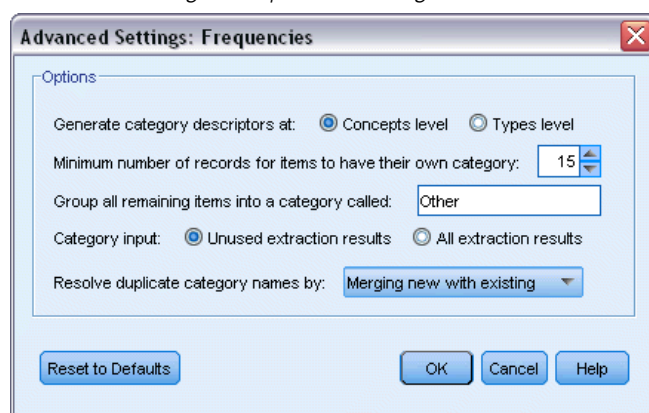
You can build categories based on a straightforward and mechanical frequency technique. With this technique, you can build one category for each item (type, concept, or pattern) that was found above a given record count. Additionally, you can build a single category for all of the less

frequently occurring items. By count, we refer to the number of records containing the extracted concept (and any of its synonyms), type, or pattern in question as opposed to the total number of occurrences in the entire text.

Grouping frequently occurring items can yield interesting results, since it may indicate a common or significant response. The technique is very useful on the unused extraction results after other techniques have been applied. Another application is to run this technique immediately after extraction when no other categories exist, edit the results to delete uninteresting categories, and then extend those categories so that they match even more records. For more information, see the topic “Extending Categories” on p. 120.

Instead of using this technique, you could sort the concepts or concept patterns by descending number of records in the Extraction Results pane and then drag and drop the top ones into the Categories pane to create the corresponding categories.

Figure 6-12
Advanced Settings: Frequencies dialog box



Generate category descriptors at. Select the kind of input for descriptors. For more information, see the topic “Building Categories” on p. 105.

- **Concepts level.** Selecting this option means that concepts or concept patterns frequencies will be used. Concepts will be used if types were selected as input for category building and concept patterns are used, if type patterns were selected. In general, applying this technique to the concept level will produce more specific results, since concepts and concept patterns represent a lower level of measurement.
- **Types level.** Selecting this option means that type or type patterns frequencies will be used. Types will be used if types were selected as input for category building and type patterns are used, if type patterns were selected. Applying this technique to the type level allows you to obtain a quick view regarding the broad range of responses given.

Minimum record count for items to have their own category. This option allows you to build categories from frequently occurring items. This option restricts the output to only those categories containing a descriptor that occurred in at least X number of records, where X is the value to enter for this option.

Group all remaining items into a category called. This option allows you to group all concepts or types occurring infrequently into a single ‘catch-all’ category with the name of your choice. By default, this category is named *Other*.

Category input. Select the group to which to apply the techniques:

- Unused extraction results. This option enables categories to be built from extraction results that are not used in any existing categories. This minimizes the tendency for records to match multiple categories and limits the number of categories produced.
- All extraction results. This option enables categories to be built using any of the extraction results. This is most useful when no or few categories already exist.

Resolve duplicate category names by. Select how to handle any new categories whose names would be the same as existing categories. You can either merge the new ones (and their descriptors) with the existing categories with the same name. Alternatively, you can choose to skip the creation of any categories if a duplicate name is found in the existing categories.

Extending Categories

Extending is a process through which descriptors are added or enhanced automatically to ‘grow’ existing categories. The objective is to produce a better category that captures related records that were not originally assigned to that category.

The automatic grouping techniques you select will attempt to identify concepts, TLA patterns, and category rules related to existing category descriptors. These new concepts, patterns, and category rules are then added as new descriptors or added to existing descriptors. The grouping techniques for extending include *concept root derivation*, *concept inclusion*, *semantic networks* (English only), and *co-occurrence rules*. The Extend empty categories with descriptors generated from the category name method generates descriptors using the words in the category names, therefore, the more descriptive the category names, the better the results.

Note: The frequency techniques are not available when extending categories.

Extending is a great way to interactively improve your categories. Here are some examples of when you might extend a category:

- After dragging/dropping concept patterns to create categories in the Categories pane
- After creating categories by hand and adding simple category rules and descriptors
- After importing a code frame in which the categories had very descriptive names
- After refining the categories that came from the TAP you chose during project creation

You can extend a category multiple times. For example, if you imported a predefined category file with very descriptive names, you could extend using the Extend empty categories with descriptors generated from the category name option to obtain a first set of descriptors, and then extend those categories again. However, in other cases, extending multiple times may result in too generic a category if the descriptors are extended wider and wider. Since the build and extend grouping techniques use similar underlying algorithms, extending directly after building categories is unlikely to produce more interesting results.

Tips:

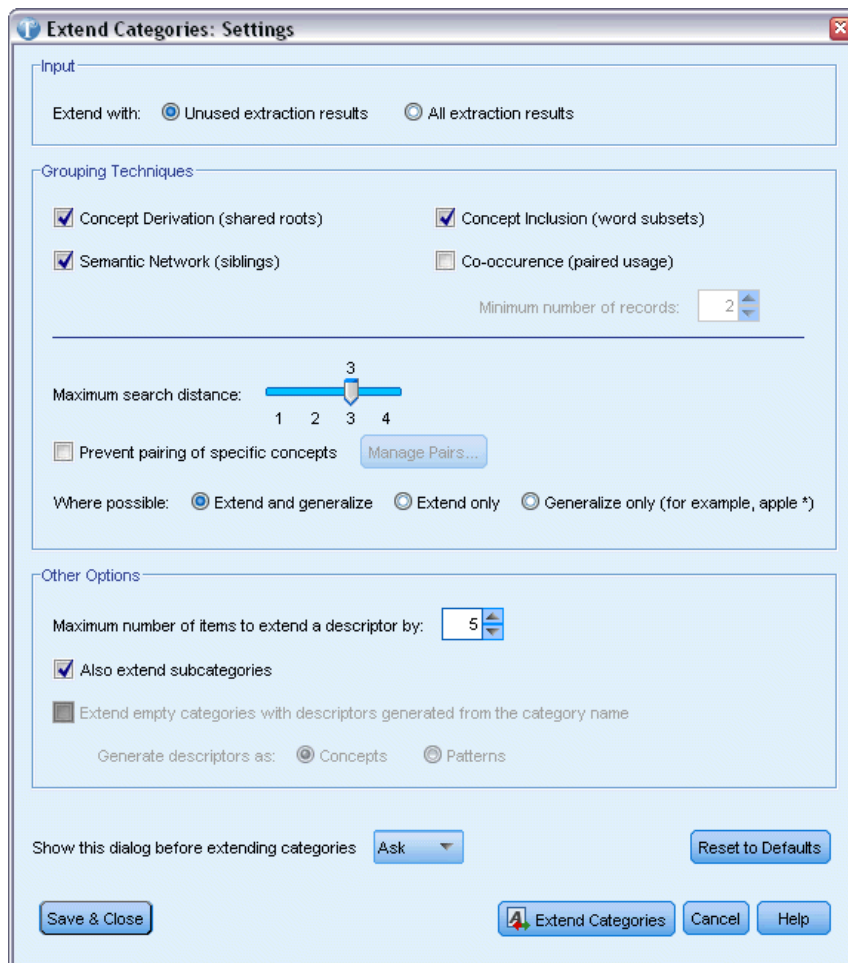
- If you attempt to extend and do not want to use the results, you can always undo the operation (Edit > Undo) immediately after having extended.
- Extending can produce two or more category rules in a category that match exactly the same set of documents since rules are built independently during the process. If desired, you can review the categories and remove redundancies by manually editing the category description. For more information, see the topic “Editing Category Descriptors” on p. 150.

To Extend Categories

- ▶ In the Categories pane, select the categories you want to extend.
- ▶ From the menus, choose Categories > Extend Categories. Unless you have chosen the option to never prompt, a message box appears.
- ▶ Choose whether you want to build now or edit the settings first.
 - Click Extend Now to begin extending categories using the current settings. The process begins and a progress dialog appears.
 - Click Edit to review and modify the settings.

After attempting to extend, any categories for which new descriptors were found are flagged by the word *Extended* in the Categories pane so that you can quickly identify them. The *Extended* text remains until you either extend again, edit the category in another way, or clear these through the context menu.

Figure 6-13
Extend Categories dialog box



Each of the techniques available when building or extending categories is well suited to certain types of data and situations, but often it is helpful to combine techniques in the same analysis to capture the full range of records. The concepts and types that were grouped into a category are still available the next time you build categories. This means that you may see a concept in multiple categories or find redundant categories.

Category Input. Select what input will be used to extend the categories:

- **Unused extraction results.** This option enables categories to be built from extraction results that are not used in any existing categories. This minimizes the tendency for records to match multiple categories and limits the number of categories produced.
- **All extraction results.** This option enables categories to be built using any of the extraction results. This is most useful when no or few categories already exist.

Grouping Techniques

For short descriptions of each of these techniques, see “Advanced Linguistic Settings” on p. 109. These techniques include:

- Concept root derivation (*not available for Japanese*)
- Semantic network (*English text only*)
- Concept inclusion
- Co-occurrence and Minimum number of docs. suboption.

A number of types are permanently excluded from the semantic networks technique since those types will not produce relevant results. They include <Positive>, <Negative>, <IP>, other non linguistic types, etc.

Maximum search distance. Select how far you want the techniques to search before producing categories. The lower the value, the fewer results you will get—however, these results will be less noisy and are more likely to be significantly linked or associated with each other. The higher the value, the more results you might get—however, these results may be less reliable or relevant. While this option is globally applied to all techniques, its effect is greatest on co-occurrences and semantic networks.

Prevent pairing of specific concepts. Select this checkbox to stop the process from grouping or pairing two concepts together in the output. To create or manage concept pairs, click Manage Pairs... For more information, see the topic “Managing Link Exception Pairs” on p. 113.

Where possible: Choose whether to simply extend, generalize the descriptors using wildcards, or both.

- **Extend and generalize.** This option will extend the selected categories and then generalize the descriptors. When you choose to generalize, the product will create generic category rules in categories using the asterisk wildcard. For example, instead of producing multiple descriptors such as [apple tart + .] and [apple sauce + .], using wildcards might produce [apple * + .]. If you generalize with wildcards, you will often get exactly the same number of records as you did before. However, this option has the advantage of reducing the number and simplifying category descriptors. Additionally, this option increases the ability to categorize more records using these categories on new text data (for example, in longitudinal/wave studies).
- **Extend only.** This option will extend your categories without generalizing. It can be helpful to first choose the Extend only option for manually-created categories and then extend the same categories again using the Extend and generalize option.
- **Generalize only.** This option will generalize the descriptors without extending your categories in any other way.

Other Options for Extending Categories

In addition to selecting the techniques to apply, you can edit any of the following options:

Maximum number of items to extend a descriptor by. When extending a descriptor with items (concepts, types, and other expressions), define the maximum number of items that can be added to a single descriptor. If you set this limit to 10, then no more than 10 additional items can be added to an existing descriptor. If there are more than 10 items to be added, the techniques stop

adding new items after the tenth is added. Doing so can make a descriptor list shorter but doesn't guarantee that the most interesting items were used first. You may prefer to cut down the size of the extension without penalizing quality by using the *Generalize with wildcards* where possible option. This option only applies to descriptors that contain the Booleans & (AND) or ! (NOT).

Also extend subcategories. This option will also extend any subcategories below the selected categories.

Extend categories with descriptors based on category names. This option attempts to automatically create descriptors for each category based on the words that make up the name of the category. The category name is scanned to see if words in the name match any extracted concepts. If a concept is recognized, it is used to find matching concept patterns and these both are used to form descriptors for the category. This option produces the best results when the category names are both long and descriptive. This is a quick method for generating category descriptors, which in turn enable the category to capture records that contain those descriptors. This option is most useful when you import categories from somewhere else or when you create categories manually with long descriptive names. This method applies only to empty categories, which have 0 descriptors. If a category already contains descriptors, it will not be extended in this way.

Generate descriptors as. This option only applies if the preceding option is selected.

- **Concepts.** Choose this option to produce the resulting descriptors in the form of concepts, regardless of whether they have been extracted from the source text.
- **Patterns.** Choose this option to produce the resulting descriptors in the form of patterns, regardless of whether the resulting patterns or any patterns have been extracted.

Creating Categories Manually

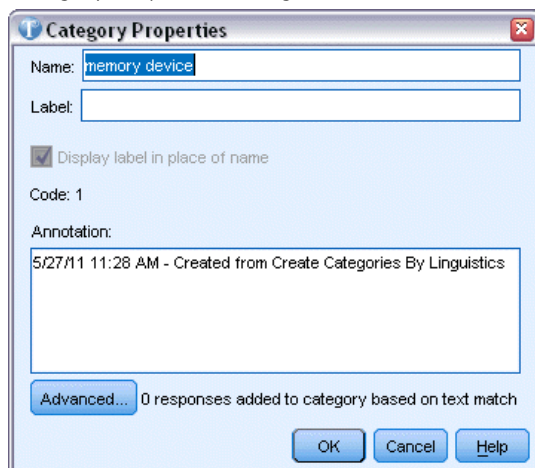
In addition to creating categories using the automated category building techniques, the Code Frame Manager, and the rule editor, you can also create categories manually. The following manual methods exist:

- Creating an empty category into which you will add elements one by one. For more information, see the topic “Creating New or Renaming Categories” on p. 124.
- Dragging terms, types, and patterns into the categories pane. For more information, see the topic “Creating Categories by Drag-and-Drop” on p. 125.

Creating New or Renaming Categories

You can create empty categories in order to add concepts and types into them. You can also rename your categories.

Figure 6-14
Category Properties dialog



To Create a New Empty Category

- ▶ Go to the Categories pane.
- ▶ From the menus, choose Categories > Create Empty Category. The dialog box opens.
- ▶ Enter a name for this category in the Name field.
- ▶ Click OK to accept the name and close the dialog box. The dialog box closes and a new category name appears in the pane.

You can now begin adding to this category. For more information, see the topic “Adding Descriptors to Categories” on p. 150.

To Rename a Category

- ▶ Select a category and choose Categories > Rename Category. The dialog box opens.
- ▶ Enter a new name for this category in the Name field.
- ▶ Click OK to accept the name and close the dialog box. The dialog box closes and a new category name appears in the pane.

Creating Categories by Drag-and-Drop

The drag-and-drop technique is manual and is not based on algorithms. You can create categories in the Categories pane by dragging:

- Extracted concepts, types, or patterns from the Extraction Results pane into the Categories pane.
- Extracted concepts from the Data pane into the Categories pane.
- Entire rows from the Data pane into the Categories pane. This will create a category made up of all of the extracted concepts and patterns contained in that row.

Note: The Extraction Results pane supports multiple selection to facilitate the dragging and dropping of multiple elements.

Important! You cannot drag and drop concepts from the Data pane that were not extracted from the text. If you want to force the extraction of a concept that you found in your data, you must add this concept to a type. Then run the extraction again. The new extraction results will contain the concept that you just added. You can then use it in your category. For more information, see the topic “Adding Concepts to Types” in Chapter 5 on p. 87.

To create categories using drag-and-drop:

- ▶ From the Extraction Results pane or the Data pane, select one or more concepts, patterns, types, records, or partial records.
- ▶ While holding the mouse button down, drag the element to an existing category or to the pane area to create a new category.
- ▶ When you have reached the area where you would like to drop the element, release the mouse button. The element is added to the Categories pane. The categories that were modified appear with a special background color. This color is called the category feedback background. For more information, see the topic “Setting Options” in Chapter 2 on p. 16.

Note: The resulting category was automatically named. If you want to change a name, you can rename it. For more information, see the topic “Editing Category Properties” on p. 149.

If you want to see which records are assigned to a category, select that category in the Categories pane. The data pane is automatically refreshed and displays all of the records for that category. To see the entire set of responses for a question, select the All Records node at the top of the category tree.

Importing and Exporting Predefined Categories

If you have your own categories stored in an Microsoft Excel (*.xls, *.xlsx) file, you can import them into IBM® SPSS® Text Analytics for Surveys.

You can also export the categories you have in an open project out to an Microsoft Excel (*.xls, *.xlsx) file. When you export your categories, you can choose to include or exclude some additional information such as descriptors and scores. For more information, see the topic “Exporting Categories” on p. 135.

Important: From SPSS Text Analytics for Surveys Version 4.0.1, predefined categories have mostly replaced the use of code frames. For example, the Import Code Frame wizard has been replaced by the Import Predefined Categories wizard. However, this new wizard still enables you to import any existing code frames you have. In addition, the Code Frame Manager is no longer supported; to edit code values select Show > Category Codes from the menus to display the Code column in the Category pane and edit any codes as required.

If your predefined categories do not have codes or you want new codes, you can automatically generate a new set of codes for the set of categories in the categories pane by choosing Categories > Manage Categories > Autogenerate Codes from the menus. This will remove any existing codes and renumber them all automatically.

Importing Predefined Categories

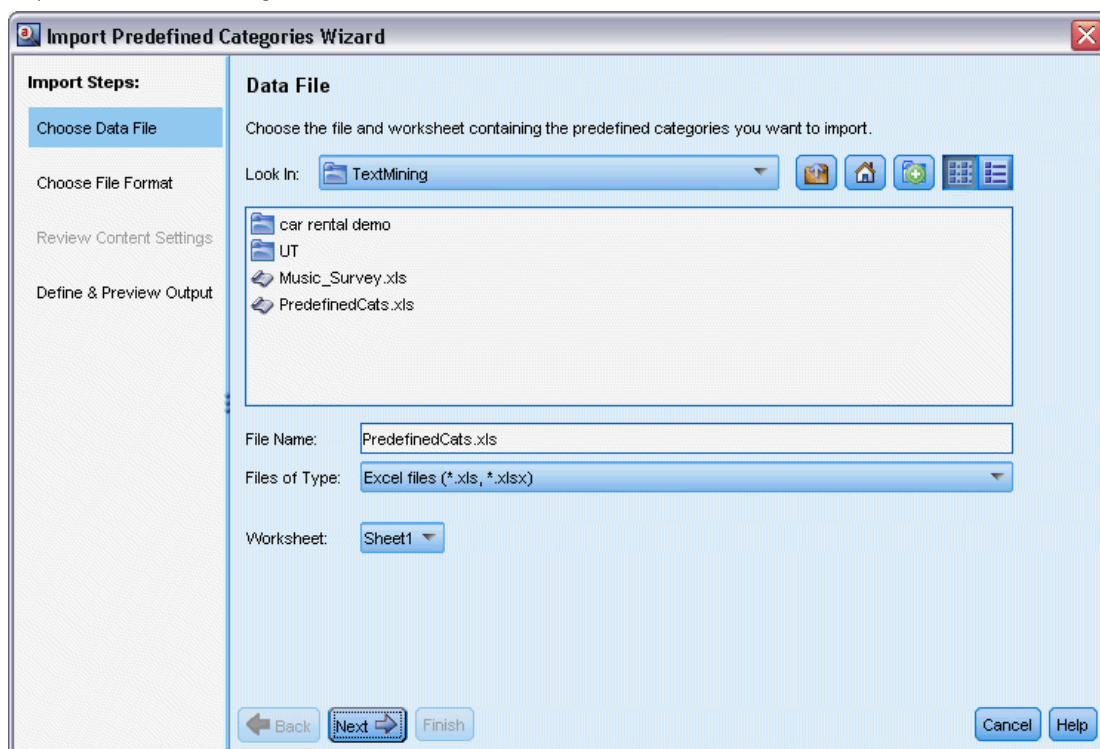
You can import your predefined categories into IBM® SPSS® Text Analytics for Surveys. Before importing, make sure the predefined category file is in an Microsoft Excel (*.xls, *.xlsx) file and is structured in one of the supportive formats. You can also choose to have the product automatically detect the format for you. The following formats are supported:

- Flat list format: For more information, see the topic “Flat List Format” on p. 131.
- Compact format: For more information, see the topic “Compact Format” on p. 132.
- Indented format: For more information, see the topic “Indented Format” on p. 133.

To Import Predefined Categories

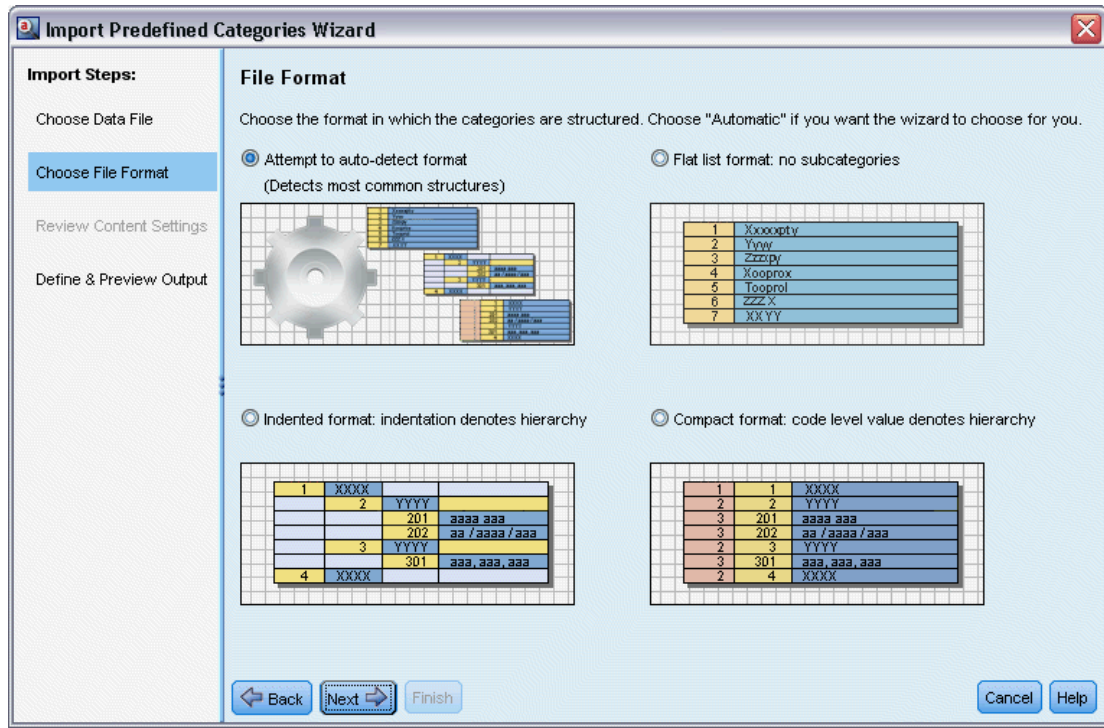
- ▶ From the menus, choose Categories > Manage Categories > Import Predefined Categories. An Import Predefined Categories wizard appears.

Figure 6-15
Import Predefined Categories wizard



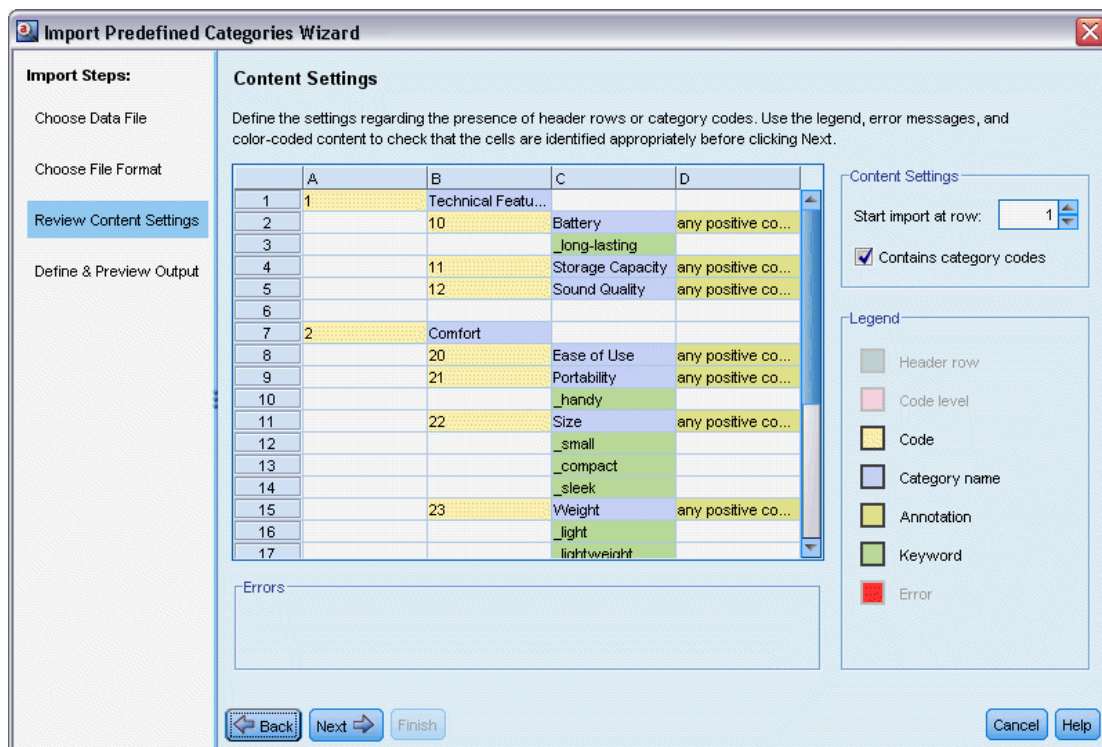
- ▶ From the Look In drop-down list, select the drive and folder in which the file is located.
- ▶ Select the file from the list. The name of the file appears in the File Name text box.
- ▶ Select the worksheet containing the predefined categories from the list. The worksheet name appears in the Worksheet field.
- ▶ Click Next to begin choosing the data format.

Figure 6-16
 Import Predefined Categories dialog box, Data Format step



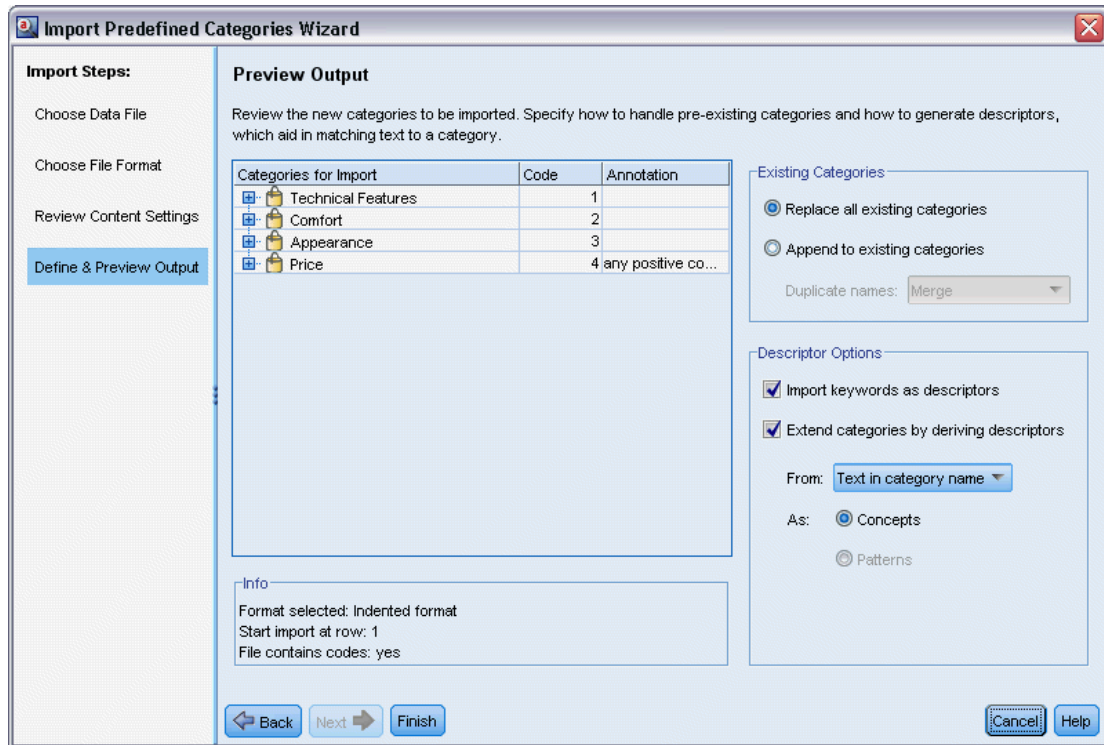
- ▶ Choose the format for your file or choose the option to allow the product to attempt to automatically detect the format. The autodetection works best on the most common formats.
 - Flat list format: For more information, see the topic “Flat List Format” on p. 131.
 - Compact format: For more information, see the topic “Compact Format” on p. 132.
 - Indented format: For more information, see the topic “Indented Format” on p. 133.
- ▶ Click Next to define the additional import options. If you choose to have the format automatically detected, you are directed to the final step.

Figure 6-17
 Import Predefined Categories, Import Options step



- ▶ If one or more rows contain column headers or other extraneous information, select the row number from which you want to start importing in the Start import at row option. For example, if your category names begin on row 7, you must enter the number 7 for this option in order to import the file correctly.
- ▶ If your file contains category codes, choose the option Contains category codes. Doing so helps the wizard properly recognize your data.
- ▶ Review the color-coded cells and legend to make sure that the data has been correctly identified. Any errors detected in the file are shown in red and referenced below the format preview table. If the wrong format was selected, go back and choose another one. If you need to make corrections to your file, make those changes and restart the wizard by selecting the file again. You must correct all errors before you can finish the wizard.
- ▶ Click Next to review the set of categories and subcategories that will be imported and to define how to create descriptors for these categories.

Figure 6-18
 Import Predefined Categories dialog box, Preview step



- ▶ Review the set of categories that will be imported in the table. If you do not see the keywords you expected to see as descriptors, it may be that they were not recognized during the import. Make sure they are properly prefixed and appear in the correct cell.
- ▶ Choose how you want to handle any pre-existing categories in your project.
 - Replace all existing categories. This option purges all existing categories and then the newly imported categories are used alone in their place.
 - Append to existing categories. This option will import the categories and merge any common categories with the existing categories. When adding to existing categories, you need to determine how you want any duplicates handled. One choice (option: Merge) is to merge any categories being imported with existing categories if they share a category name. Another choice (option: Exclude from import) is to prohibit the import of categories if one with the same name exists.
- ▶ Import keywords as descriptors is an option to import the keywords identified in your data as descriptors for the associated category.
- ▶ Extend categories by deriving descriptors is an option that will generate descriptors from the words that represent the name of the category, or subcategory, and/or the words that make up the annotation. If the words match extracted results, then those are added as descriptors to the category. This option produces the best results when the category names or annotations are both long and descriptive. This is a quick method for generating the category descriptors that enable the category to capture records that contain those descriptors.

- From field allows you to select from what text the descriptors will be derived, the names or categories and subcategories, the words in the annotations, or both.
 - As field allows you to choose to create these descriptors in the form of concepts or TLA patterns. If TLA extraction has not taken place, the options of patterns are disabled in this wizard.
- Click Finish to import the predefined categories into the Categories pane.

Flat List Format

In this flat list format, there is only one top level of categories without any hierarchy, meaning no subcategories or subnets. Category names are in a single column.

Figure 6-19
Flat List Format Example

	A	B	C
1		Technical Features	
2		_reliable	
3		_durably constructed	
4	10	Technical Features/Battery	any positive comment about long battery life
5		_long-lasting	
6	11	Technical Features/Storage Capacity	any positive comment about the amount that can be stored or memory capacity
7	12	Technical Features/Sound Quality	any positive comment about sound, quality or music quality
8	20	Comfort/Ease of Use	any positive comment indicating that it is convenient, easy and user-friendly
9	21	Comfort/Portability	any positive comment about mobility or indicating that it is handy and easy to transport
10	22	Comfort/Size	any positive comment indicating that it is small or compact
11	23	Comfort/Weight	any positive comment indicating that it is lightweight
12		_light	
13	30	Appearance/Design	any positive comment about appealing style
14		_good-looking	
15		_stylish	
16		_sleek	
17		_well-designed	
18	31	Appearance/Color	any positive comment about color

The following information can be contained in a file of this format:

- Optional **codes** column contains numerical values that uniquely identify each category. If you specify that the data file does contain codes (Contains category codes option in the Content Settings step), then a column containing unique codes for each category must exist in the cell directly to the left of category name. If your data does not contain codes, but you want to create some codes later, you can always generate codes later (Categories > Manage Categories > Autogenerate Codes). You can edit codes later by choosing Show > Category Code; the codes are displayed in a Code column in the Category pane where you can manually alter them.
- A *required* **category names** column contains all of the names of the categories. This column is required to import using this format.

- Optional **annotations** in the cell immediately to the right of the category name. This annotation consists of text that describes your categories/subcategories.
- Optional **keywords** can be imported as descriptors for categories. In order to be recognized, these keywords must exist in the cell directly below the associated category/subcategory name and the list of keywords must be prefixed by the underscore (`_`) character such as `_firearms`, `weapons / guns`. The keyword cell can contain one or more words used to describe each category. These words will be imported as descriptors or ignored depending on what you specify in the last step of the wizard. Later, descriptors are compared to the extracted results from the text. If a match is found, then that record or document is scored into the category containing this descriptor.

Table 6-3
Flat list format with codes, keywords, and annotations

Column A	Column B	Column C
Category code (<i>optional</i>)	Category name	Annotation
	<code>_Descriptor/keyword list (<i>optional</i>)</code>	

Compact Format

The compact format is structured similarly to the flat list format except that the compact format is used with hierarchical categories. Therefore, a code level column is required to define the hierarchical level of each category and subcategory.

Figure 6-20
Example of a compact predefined category file in Microsoft Excel

	A	B	C	D	E
1	1	1	1	Technical Features	
2				_reliable	
3				_durably constructed	
4	2	2	10	Battery	any positive comment about long battery life
5				_long-lasting	
6	2	2	11	Storage Capacity	any positive comment about the amount that can be stored or memory capacity
7	2	2	12	Sound Quality	any positive comment about sound, quality or music quality
8	1	1	2	Comfort	
9	2	2	20	Ease of Use	any positive comment indicating that it is convenient, easy and user-friendly
10	2	2	21	Portability	any positive comment about mobility or indicating that it is handy and easy to transport
11	2	2	22	Size	any positive comment indicating that it is small or compact
12	2	2	23	Weight	any positive comment indicating that it is lightweight
13				_light	
14	1	1	3	Appearance	
15	2	2	30	Design	any positive comment about appealing style
16				_good-looking	
17				_stylish	
18				_sleek	

The following information can be contained in a file of this format:

- A **required code level** column contains numbers that indicate the hierarchical position for the subsequent information in that row. For example, if values 1, 2 or 3 are specified and you have both categories and subcategories, then 1 is for categories, 2 is for subcategories, and 3 is for

sub-subcategories. If you have only categories and subcategories, then 1 is for categories and 2 is for subcategories. And so on, until the desired category depth.

- Optional **codes** column contains values that uniquely identify each category. If you specify that the data file does contain codes (Contains category codes option in the Content Settings step), then a column containing unique codes for each category must exist in the cell directly to the left of category name. If your data does not contain codes, but you want to create some codes later, you can always generate codes later (Categories > Manage Categories > Autogenerate Codes). You can edit codes later by choosing Show > Category Code; the codes are displayed in a Code column in the Category pane where you can manually alter them.
- A *required* **category names** column contains all of the names of the categories and subcategories. This column is required to import using this format.
- Optional **annotations** in the cell immediately to the right of the category name. This annotation consists of text that describes your categories/subcategories.
- Optional **keywords** can be imported as descriptors for categories. In order to be recognized, these keywords must exist in the cell directly below the associated category/subcategory name and the list of keywords must be prefixed by the underscore (`_`) character such as `_firearms`, `weapons / guns`. The keyword cell can contain one or more words used to describe each category. These words will be imported as descriptors or ignored depending on what you specify in the last step of the wizard. Later, descriptors are compared to the extracted results from the text. If a match is found, then that record or document is scored into the category containing this descriptor.

Table 6-4
Compact format example with codes

Column A	Column B	Column C
Hierarchical code level	Category code (<i>optional</i>)	Category name
Hierarchical code level	Subcategory code (<i>optional</i>)	Subcategory name

Table 6-5
Compact format example without codes

Column A	Column B
Hierarchical code level	Category name
Hierarchical code level	Subcategory name

Indented Format

In the Indented file format, the content is hierarchical, which means it contains categories and one or more levels of subcategories. Furthermore, its structure is indented to denote this hierarchy. Each row in the file contains either a category or subcategory, but subcategories are indented from the categories and any sub-subcategories are indented from the subcategories, and so on. You can manually create this structure in Microsoft Excel or use one that was exported from another product and saved into an Microsoft Excel format.

Figure 6-21
Example of an indented categories in Microsoft Excel

	A	B	C	D
1	1	Technical Features		
2		_reliable		
3		_durably constructed		
4			10 Battery	any positive comment about long battery life
5			_long-lasting	
6			11 Storage Capacity	any positive comment about the amount that can be stored or memory capacity
7			12 Sound Quality	any positive comment about sound, quality or music quality
8				
9	2	Comfort		
10			20 Ease of Use	any positive comment indicating that it is convenient, easy and user-friendly
11			21 Portability	any positive comment about mobility or indicating that it is handy and easy to transport
12			22 Size	any positive comment indicating that it is small or compact
13			23 Weight	any positive comment indicating that it is lightweight
14			_light	
15	3	Appearance		
16			30 Design	any positive comment about appealing style
17			_good-looking	
18			_stylish	

- **Top level category codes and category names** occupy the columns A and B, respectively. Or, if no codes are present, then the category name is in column A.
- **Subcategory codes and subcategory names** occupy the columns B and C, respectively. Or, if no codes are present, then the subcategory name is in column B. The subcategory is a member of a category. You cannot have subcategories if you do not have top level categories.

Table 6-6
Indented structure with codes

Column A	Column B	Column C	Column D
Category code (optional)	Category name		
	Subcategory code (optional)	Subcategory name	
		Sub-subcategory code (optional)	Sub-subcategory name

Table 6-7
Indented structure without codes

Column A	Column B	Column C
Category name		
	Subcategory name	
		Sub-subcategory name

The following information can be contained in a file of this format:

- Optional **codes** must be values that uniquely identify each category or subcategory. If you specify that the data file does contain codes (Contains category codes option in the Content Settings step), then a unique code for each category or subcategory must exist in the cell directly to the left of category/subcategory name. If your data does not contain codes, but

you want to create some codes later, you can always generate codes later (Categories > Manage Categories > Autogenerate Codes). You can edit codes later by choosing Show > Category Code; the codes are displayed in a Code column in the Category pane where you can manually alter them.

- A *required name* for each category and subcategory. Subcategories must be indented from categories by one cell to the right in a separate row.
- Optional **annotations** in the cell immediately to the right of the category name. This annotation consists of text that describes your categories/subcategories.
- Optional **keywords** can be imported as descriptors for categories. In order to be recognized, these keywords must exist in the cell directly below the associated category/subcategory name and the list of keywords must be prefixed by the underscore (`_`) character such as `_firearms`, `weapons / guns`. The keyword cell can contain one or more words used to describe each category. These words will be imported as descriptors or ignored depending on what you specify in the last step of the wizard. Later, descriptors are compared to the extracted results from the text. If a match is found, then that record or document is scored into the category containing this descriptor.

Important! If you use a code at one level, you must include a code for each category and subcategory. Otherwise, the import process will fail.

Exporting Categories

You can also export the categories you have in an open project into an Microsoft Excel (*.xls, *.xlsx) file format. The data that will be exported comes largely from the current contents of the Categories pane or from the category properties. Therefore, we recommend that you score again if you plan to also export the Docs. score value.

Always gets exported...

- Category codes, if present
- Category (and subcategory) names
- Code levels, if present (*Flat/Compact* format)
- Column headings (*Flat/Compact* format)

Exported optionally...

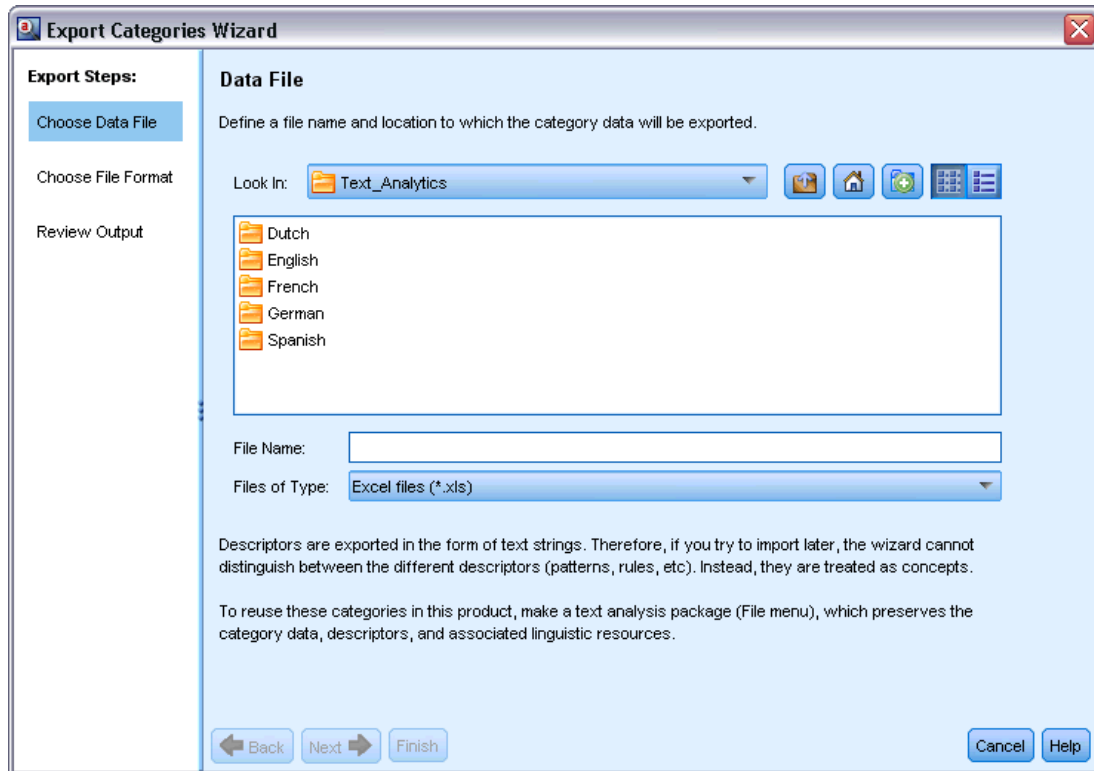
- Docs. scores
- Category annotations
- Descriptor names
- Descriptors counts

Important! When you export descriptors, they are converted to text strings and prefixed by an underscore. If you re-import into this product, the ability to distinguish between descriptors that are patterns, those that are category rules, and those that are plain concepts is lost. If you intend to reuse these categories in this product, we highly recommend making a text analysis package (TAP) file instead since the TAP format will preserve all descriptors as they are currently defined as well as all your categories, codes, and also the linguistic resources used. TAP files can be used in both IBM® SPSS® Text Analytics and IBM® SPSS® Text Analytics for Surveys. For more information, see the topic “Using Text Analysis Packages” in Chapter 3 on p. 40.

To Export Predefined Categories

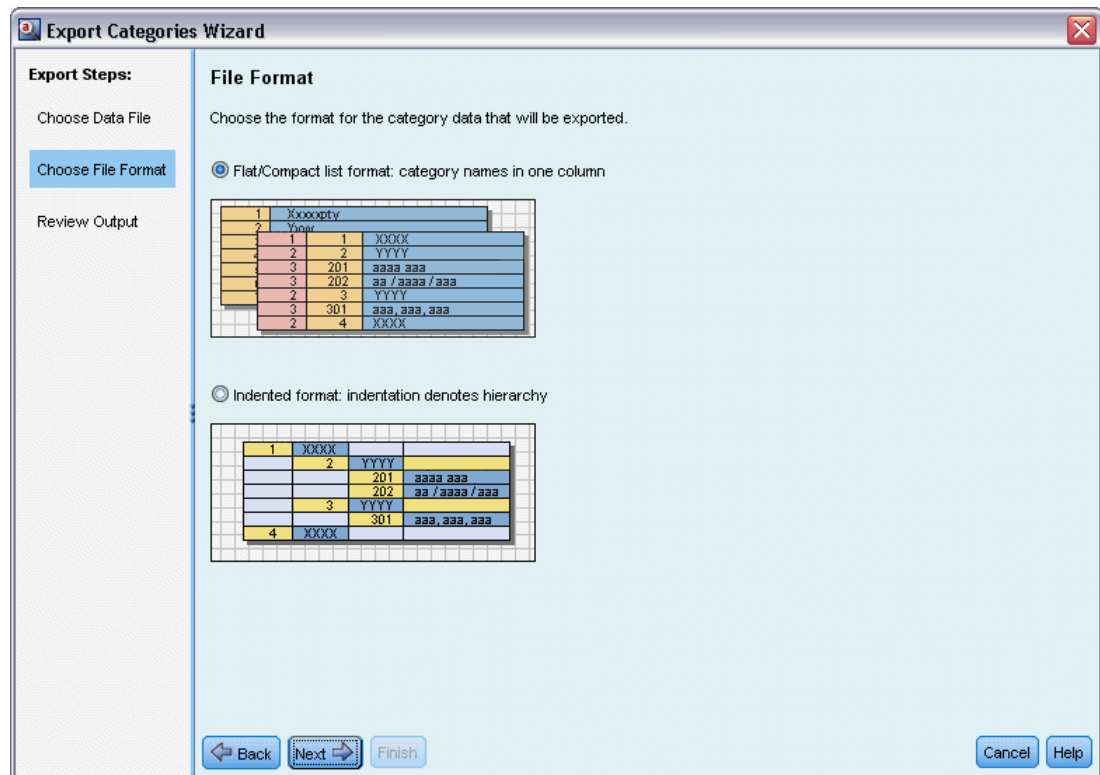
- ▶ From the menus, choose Categories > Manage Categories > Export Categories. An Export Categories wizard appears.

Figure 6-22
Export Categories wizard, step 1



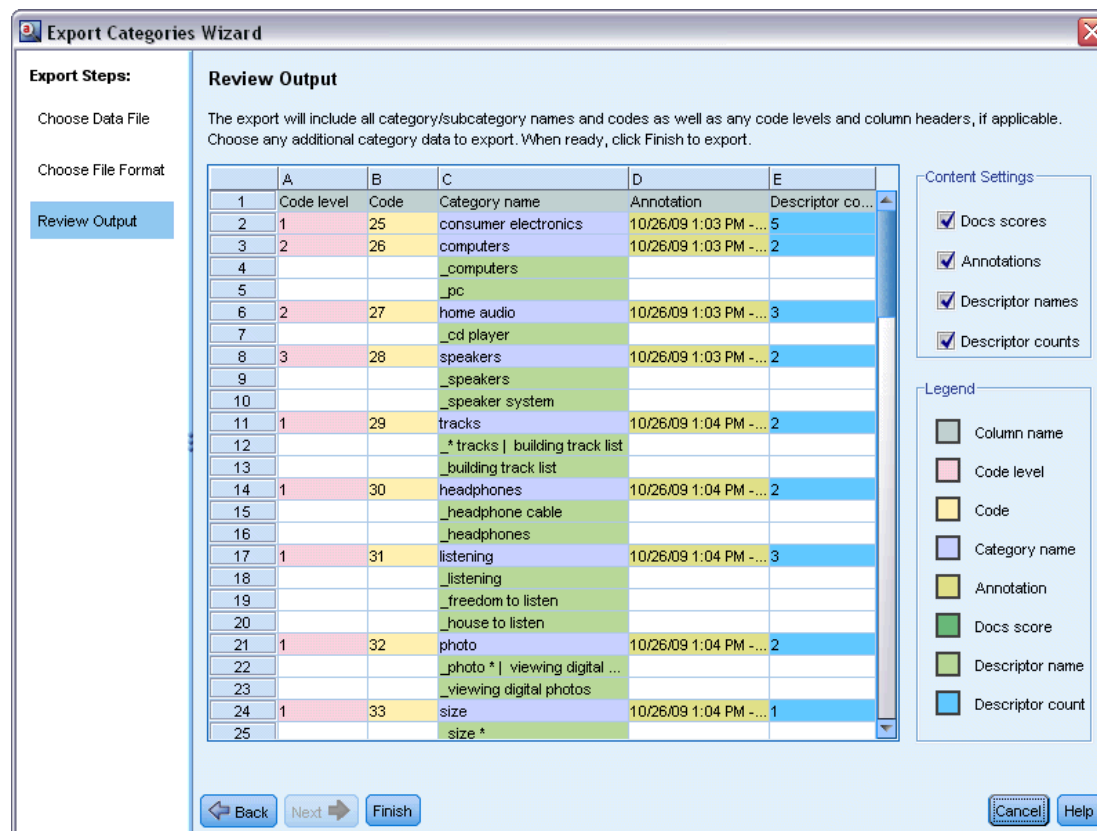
- ▶ Choose the location and enter the name of the file that will be exported.
- ▶ Enter a name for the output file in the File Name text box.
- ▶ Click Next to choose the format into which you will export your category data.

Figure 6-23
Export Categories wizard, step 2



- ▶ Choose the format from the following:
 - Flat or Compact list format: For more information, see the topic “Flat List Format” on p. 131. Flat list contains no subcategories. For more information, see the topic “Compact Format” on p. 132. Compact list format contains hierarchical categories.
 - Indented format: For more information, see the topic “Indented Format” on p. 133.
- ▶ Click Next to begin choosing the content to be exported and to review the proposed data.

Figure 6-24
Export Categories wizard, step 3



- ▶ Review the content for the exported file.
- ▶ Select or unselect the additional content settings to be exported such as Annotations or Descriptor names.
- ▶ Click Finish to export the categories.

Using Category Rules

You can create categories in many ways. One of these ways is to define category rules to express ideas. Category rules are statements that automatically classify records into a category based on a logical expression using extracted concepts, types, and patterns as well as Boolean operators. For example, you could write an expression that means *include all records that contain the extracted concept embassy but not argentina in this category*.

While some category rules are produced automatically when building categories using grouping techniques such as *co-occurrence* and *concept root derivation* (Categories > Build Settings > Advanced Settings: Linguistics), you can also create category rules manually in the rule editor using your category understanding of the data and context. Each rule is attached to a single category so that each record matching the rule is then scored into that category.

Category rules help enhance the quality and productivity of your text mining results and further quantitative analysis by allowing you to categorize responses with greater specificity. Your experience and business knowledge might provide you with a specific understanding of your data and context. You can leverage this understanding to translate that knowledge into category rules to categorize your records even more efficiently and accurately by combining extracted elements with Boolean logic.

The ability to create these rules enhances coding precision, efficiency, and productivity by allowing you to layer your business knowledge onto the product's extraction technology.

Note: For examples of how rules match text, see “Category Rule Examples” on p. 144

Category Rule Syntax

While some category rules are produced automatically when building categories using grouping techniques such as *co-occurrence* and *concept root derivation* (Categories > Build Settings > Advanced Settings: Linguistics), you can also create category rules manually in the rule editor. Each rule is a descriptor of a single category; therefore, each record matching the rule is automatically scored into that category.

Note: For examples of how rules match text, see “Category Rule Examples” on p. 144

When you are creating or editing a rule, you must have it open in the rule editor. You can add concepts, types, or patterns as well as use wildcards to extend the matches. When you use extracted concepts, types and patterns, you can benefit from finding all related concepts.

Important! To avoid common errors, we recommend dragging and dropping concepts directly from the Extraction Results pane, Text Link Analysis panes, or the Data pane into the rule editor or adding them via the context menus whenever possible.

When concepts, types, and patterns are recognized, an icon appears next to the text.



Extracted concept



Extracted type



Extracted pattern

Rule Syntax and Operators

The following table contains the characters with which you'll define your rule syntax. Use these characters along with the concepts, types, and patterns to create your rule.

Table 6-8
Supported syntax

Character	Description
&	The “and” boolean. For example, a & b contains both a <i>and</i> b such as: - invasion & united states - 2016 & olympics - good & apple
	The “or” boolean is inclusive, which means that if any or all of the elements are found, a match is made. For example, a b contains either a <i>or</i> b such as: - attack france - condominium apartment
! ()	The “not” boolean. For example, !(a) does not contain a. such as, !(good & hotel)

Character	Description
*	A wildcard representing anything from a single character to a whole word depending how it is used. For more information, see the topic “Using Wildcards in Category Rules” on p. 142.
()	An expression delimiter. Any expression within the parenthesis is evaluated first.
+	The pattern connector used to form an order-specific pattern. When present, the square brackets must be used. For more information, see the topic “Using TLA Patterns in Category Rules” on p. 140.
[]	The pattern delimiter is required if you are looking to match based on an extracted TLA pattern inside of a category rule. The content within the brackets refers to TLA patterns and will never match concepts or types based on simple co-occurrence. If you did not extract this TLA pattern, then no match will be possible. For more information, see the topic “Using TLA Patterns in Category Rules” on p. 140. Do not use square brackets if you are looking to match concepts and types instead of patterns. <i>Note:</i> In older versions, co-occurrence and synonym rules generated by the category building techniques used to be surrounded by square brackets. In all new versions, square brackets now indicate the presence of a TLA pattern. Instead, rules produced by the co-occurrence technique and synonyms will be encapsulated in parentheses, such as (speaker systems speakers).

The & and | operators are commutative such that $a \& b = b \& a$ and $a | b = b | a$.

Escaping Characters with Backslash

If you have a concept that contains any character that is also a syntax character you must place a backslash in front of that character so that the rule is properly interpreted. The backslash (\) character is used to escape characters that otherwise have a special meaning. When you drag and drop into the editor, backslashing is done for you automatically.

The following rule syntax characters must be preceded by a backslash if you want it treated as it is rather than as rule syntax:

```
&!|+<>()[]*
```

For example, since the concept `r&d` contains the “and” operator (&), the backslash is required when it is typed into the rule editor, such as: `r\&d`.

Using TLA Patterns in Category Rules

Text link analysis patterns can be explicitly defined in category rules to allow you to obtain even more specific and contextual results. When you define a pattern in a category rule, you are bypassing the more simple concept extraction results and only matching documents and records based on extracted text link analysis pattern results.

Delimiting with square brackets. A TLA pattern must be surrounded by square brackets [] if you are using it inside of a category rule. The pattern delimiter is required if you are looking to match based on an extracted TLA pattern. Since category rules can contain, types, concepts, or patterns, the brackets clarify to the rule that the contents within the brackets refers to extracted TLA pattern. If you did not extract this TLA pattern, then no match will be possible. If you see a pattern without brackets such as `apple + good` in the Categories pane, this likely means that the pattern was added directly to the category outside of the category rule editor. For example, if you add a concept pattern directly to category from extraction results pane, it will not appear with

square brackets. However, when using a pattern within a category rule, you must encapsulate the pattern within the square brackets inside the category rule such as [banana + !(good)].

Using the + sign in patterns. In IBM® SPSS® Text Analytics for Surveys, you can have two-part patterns. When you want to indicate that the order is important, use the + sign to connect the elements. If order is unimportant, you can use the & boolean instead. In the following two texts, the position of the word *better* is important: “*broccoli is better*” and “*it’s better than broccoli*”.

For example, let’s say you had the two following sample texts the expression: “*I like pineapple*” and “*I hate pineapple. However, I like strawberries*”. The expression like & pineapple would match both texts as it is a concept expression and not a text link rule (not enclosed in brackets). The expression pineapple + like matches only “*I like pineapple*” since in the second text, the word *like* is associated to *strawberries* instead.

Grouping with patterns. You can simplify your rules with your own patterns. Let’s say you want to capture the following three expressions, cayenne peppers + like, chili peppers + like, and peppers + like. You can group them into a single category rule such as [*peppers & like]. If you had another expression hot peppers + good, you can group those four with a rule such as [*peppers + <Positive>].

Order in patterns. In order to better organize output, the text link analysis rules supplied in the templates you installed with your product attempt to output basic patterns in the same order regardless of word order in the sentence. For example, if you had a record containing the text, “*Good presentations.*” and another record containing “*the presentations were good*”, both text are matched by the same rule and output in the same order as presentation + good in the concept pattern results rather than presentation + good and also good + presentation. And in two-slot pattern such as those in the example, the concepts assigned to types in the Opinions library will be presented last in the output by default such as apple + bad.

Table 6-9
Pattern syntax and boolean usage

Expression	Matches a record that
[]	Contains any TLA pattern. The pattern delimiter is required <i>in category rules</i> if you are looking to match based on an extracted TLA pattern. The content within the brackets refers to TLA patterns not simple concepts and types. If you did not extract this TLA pattern, then no match will be possible. If you wanted to create a rule that did not include any patterns, you could use !([]).
[a]	Contains a pattern of which at least one element is a regardless of its position in the pattern. For example, [deal] can match [deal + good] or just [deal + .]
[a + b]	Contains a concept pattern. For example, [deal + good]. <i>Note:</i> If you only want to capture this pattern without adding any other elements, we recommend adding the pattern directly to your category rather than making a rule with it.
[<A> +]	Contains any pattern with type <A> in the first slot and type in the second slot, and there are exactly two slots. The + sign denotes that the order of the matching elements is important. For example, [<Budget> + <Negative>]. <i>Note:</i> If you only want to capture this pattern without adding any other elements, we recommend adding the pattern directly to your category rather than making a rule with it.

Expression	Matches a record that
[<A> &]	Contains any type pattern with type <A> and type . For example, [<Budget> & <Negative>]. This TLA pattern will never be extracted; however, when written as such it is really equal to [<Budget> + <Negative>] [<Negative> + <Budget>]. The order of the matching elements is unimportant.
[a + .]	Contains a pattern where a is the only concept and there is nothing in any other slots for that pattern. For example, [deal + .] matches the concept pattern where the only output is the concept deal. If you added the concept deal as a category descriptor, you would get all records with deal as a concept including positive statements about a deal. However, using [deal + .] will match only those records pattern results representing deal and no other relationships or opinions and would not match deal + fantastic. <i>Note:</i> If you only want to capture this pattern without adding any other elements, we recommend adding the pattern directly to your category rather than making a rule with it.
[<A> + <>]	Contains a pattern where <A> is the only type. For example, [<Budget> + <>] matches the pattern where the only output is a concept of the type <Budget>. <i>Note:</i> You can use the <> to denote an empty type only when putting it after the pattern + symbol in type pattern such as [<Budget> + <>] but not [price + <>]. <i>Note:</i> If you only want to capture this pattern without adding any other elements, we recommend adding the pattern directly to your category rather than making a rule with it.
[a + !(b)]	Contains at least one pattern that includes the concept a but does not include the concept b. Must include at least one pattern. For example, [price + !(high)] or for types, [!(<Fruit> <Vegetable>) + <Positive>]
!([<A> &])	Does not contain a specific pattern. For example, !([<Budget> & <Negative>]).

Note: For examples of how rules match text, see “Category Rule Examples” on p. 144

Using Wildcards in Category Rules

Wildcards can be added to concepts in rules in order to extend the matching capabilities. The asterisk * wildcard can be placed before and/or after a word to indicate how concepts can be matched. There are two types of wildcard uses:

- **Affix wildcards.** These wildcards immediately prefix or suffix without any space separating the string and the asterisk. For example, *operat** could match *operat*, *operate*, *operates*, *operations*, *operational*, and so on.
- **Word wildcards.** These wildcards prefix or suffix a concept with a space between the concept and the asterisk. For example, ** operation* could match *operation*, *surgical operation*, *post operation*, and so on. Additionally, a word wildcard can be used along side an affix wildcard such as, ** operat* **, which could match *operation*, *surgical operation*, *telephone operator*, *operatic aria*, and so on. As you can see in this last example, we recommend that wildcards be used with care so as not to cast the net too widely and capture unwanted matches.

Exceptions!

- A wildcard can never stand on its own. For example, (apple | *) would not be accepted.
- A wildcard can never be used to match type names. <Negative*> will not match any type names at all.

- You cannot filter out certain types from being matched to concepts found through wildcards. The type to which the concept is assigned is used automatically.
- A wildcard can never be in the middle of a word sequence, whether it is end or beginning of a word (open* account) or a standalone component (open * account). You cannot use wildcards in type names either. For example, word* word, such as apple* recipe, will not match applesauce recipe or anything else at all. However, apple* * would match *applesauce recipe*, *apple pie*, *apple* and so on. In another example, word * word, such as apple * toast, will not match *apple cinnamon toast* or anything else at all since the asterisk appears between two other words. However, apple * would match *apple cinnamon toast*, *apple*, *apple pie* and so on.

Table 6-10
Wildcard usage

Expression	Matches a record that
*apple	Contains a concept that ends with letter written but may have any number of letters as a prefix. For example: *apple ends with the letters <i>apple</i> but can take a prefix such as: - apple - pineapple - crabapple
apple*	Contains a concept that starts with letters written but may have any number of letters as a suffix. For example: apple* starts with the letters <i>apple</i> but can take a suffix or no suffix such as: - apple - applesauce - applejack For example, apple* & !(pear* quince), which contains a concept that starts with the letters <i>apple</i> but not a concept starting with the letters <i>pear</i> or the concept <i>quince</i> , would NOT match: apple & quince but could match: - applesauce - apple & orange
product	Contains a concept that contains the letters written product, but may have any number of letters as either a prefix or suffix or both. For example: *product* could match: - product - byproduct - unproductive
* loan	Contains a concept that contains the word loan but may be a compound with another word placed before it. For example, * loan could match: - loan - car loan - home equity loan For example, [* delivery + <Negative>] contains a concept that ends in the word <i>delivery</i> in the first position and contains a type <Negative> in the second position could match the following concept patterns: - package delivery + slow - overnight delivery + late

Expression	Matches a record that
event *	Contains a concept that contains the word <code>event</code> but may be a compound followed by another word. For example, <code>event *</code> could match: <ul style="list-style-type: none"> - event - event location - event planning committee
* apple *	Contains a concept that might start with any word followed by the word <code>apple</code> possibly followed by another word. * means 0 or n, so it also matches <code>apple</code> . For example, <code>* apple *</code> could match: <ul style="list-style-type: none"> - gala applesauce - granny smith apple crumble - famous apple pie - apple For example, <code>[* reservation* * + <Positive>]</code> , which contains a concept with the word <code>reservation</code> (regardless of where it is in the concept) in the first position and contains a type <code><Positive></code> in the second position could match the concept patterns: <ul style="list-style-type: none"> - reservation system + good - online reservation + good

Note: For examples of how rules match text, see “Category Rule Examples” on p. 144

Category Rule Examples

To help demonstrate how rules are matched to records differently based on the syntax used to express them, consider the following example.

Example Records

Imagine you had two records:

- **Record A:** “*when I checked my wallet, I saw I was missing 5 dollars.*”
- **Record B:** “*\$5 was found at the picnic area, but the blanket was missing.*”

The following two tables show what might be extracted for concepts and types as well as concept patterns and type patterns.

Concepts and Types Extracted From Example

Table 6-11
Example Extracted Concepts and Types

Extracted Concept	Concepts Typed As
wallet	<Unknown>
missing	<Negative>
USD5	<Currency>
blanket	<Unknown>
picnic area	<Unknown>

TLA Patterns Extracted From Example

Table 6-12
Example Extracted TLA Pattern Output

Extracted Concept Patterns	Extracted Type Patterns	From Record
picnic area + .	<Unknown> + <>	Record B

Extracted Concept Patterns	Extracted Type Patterns	From Record
wallet + .	<Unknown> + <>	Record A
blanket + missing	<Unknown> + <Negative>	Record B
USD5 + .	<Currency> + <>	Record B
USD5 + missing	<Currency> + <Negative>	Record A

How Possible Category Rules Match

The following table contains some syntax that could be entered in the category rule editor. Not all rules here work and not all match the same records. See how the different syntax affects the records matched.

Table 6-13
Sample Rules

Rule Syntax	Result
USD5 & missing	Matches both records A and B since they both contain the extracted concept missing and the extracted concept USD5. This is equivalent to: (USD5 & missing)
missing & USD5	Matches both records A and B since they both contain the extracted concept missing and the extracted concept USD5. This is equivalent to: (missing & USD5)
missing & <Currency>	Matches both records A and B since they both contain the extracted concept missing and a concept matching the type <Currency>. This is equivalent to: (missing & <Currency>)
<Currency> & missing	Matches both records A and B since they both contain the extracted concept missing and a concept matching the type <Currency>. This is equivalent to: (<Currency> & missing)
[USD5 + missing]	Matches A but not B since record B did not produce any TLA pattern output containing USD5 + missing (see previous table). This is equivalent to the TLA pattern output: USD5 + missing
[missing + USD5]	Matches neither record A nor B since no extracted TLA pattern (see previous table) match the order expressed here with missing in the first position. This is equivalent to the TLA pattern output: USD5 + missing
[missing & USD5]	Matches A but not B since no such TLA pattern was extracted from record B. Using the character & indicates that order is unimportant when matching; therefore, this rule looks for a pattern match to either [missing + USD5] or [USD5 + missing]. Only [USD5 + missing] from record A has a match.
[missing + <Currency>]	Matches neither record A nor B since no extracted TLA pattern matched this order. This has no equivalent, since a TLA output is only based on terms (USD5 + missing) or on types (<Currency> + <Negative>), but does not mix concepts and types.

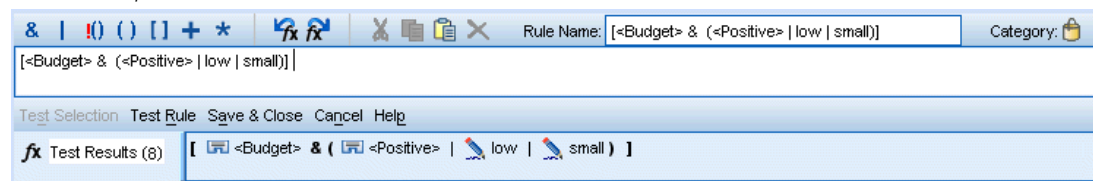
Rule Syntax	Result
[<Currency> + <Negative>]	Matches record A but not B since no TLA pattern was extracted from record B. This is equivalent to the TLA output: <Currency> + <Negative>
[<Negative> + <Currency>]	Matches neither record A nor B since no extracted TLA pattern matched this order. In the <code>Opinions</code> template, by default, when a <i>topic</i> is found with an <i>opinion</i> , the <i>topic</i> (<Currency>) occupies the first slot position and <i>opinion</i> (<Negative>) occupies the second slot position.

Creating Category Rules

When you are creating or editing a rule, you must have the rule open in the rule editor. You can add concepts, types, or patterns as well as use wildcards to extend the matches. When you use recognized concepts, types and patterns, you benefit since it will find all related concepts. For example, when you use a concept, all of its associated terms, plural forms, and synonyms are also matched to the rule. Likewise, when you use a type, all of its concepts are also captured by the rule.

You can open the rule editor by editing an existing rule or by right-clicking the category name and choosing `Create Rule`.

Figure 6-25
Rule editor pane



You can use context menus, drag-and-drop, or manually enter concepts, types, and patterns into the editor. Then combine these with Boolean operators (&, ! (), |) and brackets to form your rule expressions. To avoid common errors, we recommend dragging and dropping concepts directly from the Extraction Results pane or the Data pane into the rule editor. Pay close attention to the syntax of the rules to avoid errors. For more information, see the topic “Category Rule Syntax” on p. 139.

Note: For examples of how rules match text, see “Category Rule Examples” on p. 144

To Create a Rule

- ▶ If you have not yet extracted any data or your extraction is out of date, do so now. For more information, see the topic “Extracting Data” in Chapter 5 on p. 81.
- ▶ In the Categories pane, select the category in which you want to add your rule.
- ▶ From the menus choose `Categories > Create Rule`. The category rule editor pane opens in the window.
- ▶ In the Rule Name field, enter a name for your rule. If you do not provide a name, the expression will be used as the name automatically. You can rename this rule later.

- ▶ In the larger expression text field, you can:
 - Enter text directly in the field or drag-and-drop from another pane. Use only extracted concepts, types, and patterns. For example, if you enter the word `cats` but only the singular form, `cat`, appears in your Extraction Results pane, the editor will not be able to recognize `cats`. In this last case, the singular form might automatically include the plural, otherwise you could use a wildcard. For more information, see the topic “Category Rule Syntax” on p. 139.
 - Select the concepts, types, or patterns you want to add to rules and use the menus.
 - Add Boolean operators to link elements in your rule together. Use the toolbar buttons to add the “and” Boolean `&`, the “or” Boolean `|`, the “not” Boolean `!`, parentheses `()`, and brackets for patterns `[]` to your rule.
- ▶ Click the Test Rule button to verify that your rule is well-formed. For more information, see the topic “Category Rule Syntax” on p. 139. The number of records found appears in parentheses next to the text Test result. To the right of this text, you can see the elements in your rule that were recognized or any error messages. If the graphic next to the type, pattern, or concept appears with a red question mark, this indicates that the element does not match any known extractions. If it does not match, then the rule will not find any records.
- ▶ To test a part of your rule, select that part and click Test Selection.
- ▶ Make any necessary changes and retest your rule if you found problems.
- ▶ When finished, click Save & Close to save your rule again and close the editor. The new rule name appears in the category.

Figure 6-26
Rule in Categories pane

Category	#Descriptors	#Responses
Other: No Like-No Dislike	4	0
Pos: Product: Information	4	0
Other: No Experience-Does Not Apply	4	1
Pos: Service: Orders-Contracts	4	0
Neg: Product: Design-Features	4	3
Contx: Company: Public Image-Reputation	3	1
Pos: No Plan to Change-Would Recommend	3	0
Pos: Pricing and Billing	3	9
<fx [<Budget> & (<Positive> low small)]		1
<fx [reasonable + .]		0
<fx [<PositiveBudget>]		8
Neg: Plan to Change-Not Recommended	3	0
Contx: Service	2	0
Contx: Quality	1	1

Editing and Deleting Rules

After you have created and saved a rule, you can edit that rule at any time. For more information, see the topic “Category Rule Syntax” on p. 139.

If you no longer want a rule, you can delete it.

To Edit Rules

- ▶ In the categories pane, select the rule you want to edit.
- ▶ From the menus choose Categories > Edit Rule or double-click the rule name. The editor opens with the selected rule.
- ▶ Make any changes to the rule using extraction results and the toolbar buttons.
- ▶ Retest your rule to make sure that it returns the expected results.
- ▶ Click Save & Close to save your rule again and close the editor.

To Delete a Rule

- ▶ In the categories pane, select the rule you want to delete.
- ▶ From the menus, choose Edit > Delete. The rule is deleted from the category.

Editing and Refining Categories

Once you create some categories, you will invariably want to examine them and make some adjustments. In addition to refining the linguistic resources, you should review your categories by looking for ways to combine or clean up their definitions, as well as checking some of the categorized records. You can also review the records in a category and make adjustments so that categories are defined in such a way that nuances and distinctions are captured.

You can use the built-in, automated, category-building techniques to create your categories; however, you are likely to want to perform a few tweaks to these categories. After using one or more technique, a number of new categories appear in the window. You can then review the data in a category and make adjustments until you are comfortable with your category definitions. For more information, see the topic “About Categories” on p. 103.

Here are some options for refining your categories, most of which are described in the following pages:

- Editing category properties (renaming, adding labels, adding annotations)
- Adding descriptors to your categories
- Editing categories
- Moving categories
- Flattening hierarchical categories
- Merging categories together
- Adding text matching
- Forcing responses into categories
- Copying and reusing categories
- Deleting categories
- Making changes to your linguistic resources and reextracting
- Visualizing how your categories work together and making adjustments. For more information, see the topic “Visualizing Graphs” in Chapter 7 on p. 159.

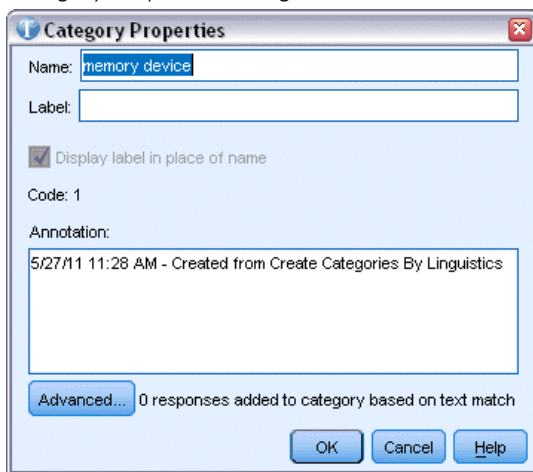
Editing Category Properties

Like many other elements in IBM® SPSS® Text Analytics for Surveys, you can edit the properties of your categories: name, label, annotations, and advanced text match entries. For more information, see the topic “Category Properties” on p. 104. In addition to the properties you can edit, you can also see the number of items included in the category definition, meaning the number of term, types, TLA patterns, or category rules that make up that category. The code number is also shown and corresponds to the code value found in the Code Frame Manager.

To Edit Category Properties:

- ▶ From the menus choose Categories > Category Properties. The Category Properties dialog box opens.

Figure 6-27
Category Properties dialog



- ▶ If desired, rename the category by entering a new name in the Name field.
- ▶ Change the category name or label.
- ▶ To use the label in the interface, such as in the Category pane, instead of the category name, select Display label in place of name.
- ▶ If desired, enter an annotation in the Annotation field.
- ▶ To force a word or phrase into the category definition, click Advanced and enter your text matches in the table. For more information, see the topic “Text Matching in Categories” on p. 154.
- ▶ Click OK to apply your changes.

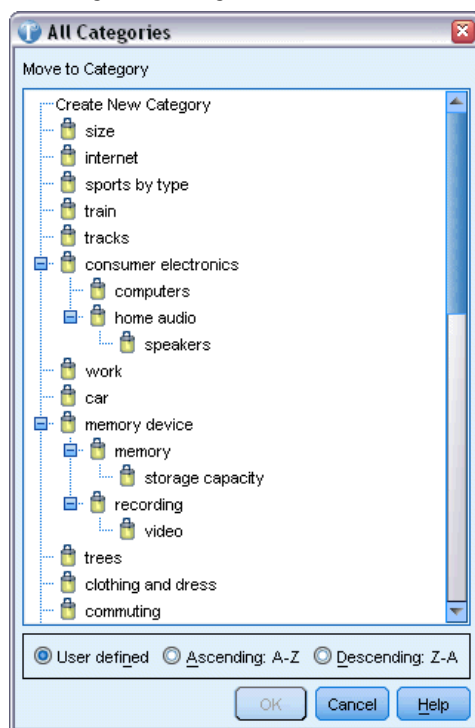
Adding Descriptors to Categories

After using automated techniques, you will most likely still have extraction results that were not used in any of the category definitions. You should review this list in the Extraction Results pane. If you find elements that you would like to move into a category, you can add them to an existing or new category.

To Add a Concept or Type to a Category

- ▶ From within the Extraction Results and Data panes, select the elements that you want to add to a new or existing category.
- ▶ From the menus, choose Categories > Add to Category. The All Categories dialog box to presents the set of categories. Select the category to which you want to add the selected elements. If you want to add the elements to a new category, select New Category. A new category appears in the Categories pane using the name of the first selected element.

Figure 6-28
All Categories dialog box



Editing Category Descriptors

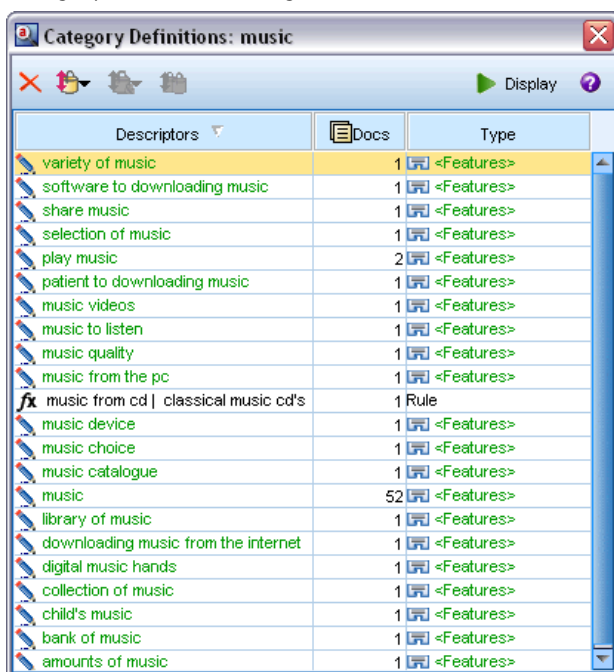
Once you have created some categories, you can open each category to see all of the descriptors that make up its definition. Inside the Category Definitions dialog box, you can make a number of edits to your category descriptors. Also, if categories are shown in the category tree, you can also work with them there.

To Edit a Category

- ▶ Select the category you want to edit in the Categories pane.
- ▶ From the menus, choose View > Category Definitions. The Category Definitions dialog box opens.

Figure 6-29

Category Definitions dialog box



- ▶ Select the descriptor you want to edit and click the corresponding toolbar button.

The following table describes each toolbar button that allows you to edit your category definitions.

Table 6-14

Toolbar buttons and descriptions

Icons	Description
	Deletes the selected descriptors from the category.
	Moves the selected descriptors to a new or existing category.
	Moves the selected descriptors in the form of an & category rule to a category. For more information, see the topic “Using Category Rules” on p. 138.
	Moves each of the selected descriptors as its own new category
Display	Updates what is displayed in the Data pane and the Visualization pane according to the selected descriptors

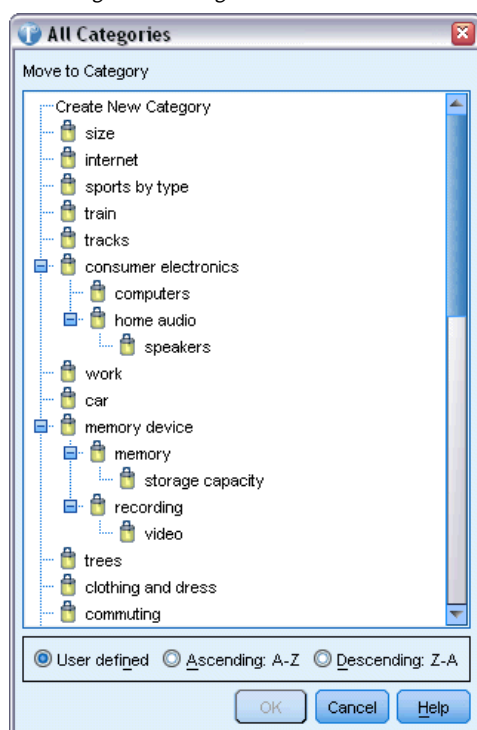
Moving Categories

If you want to place a category into another existing category or move descriptors into another category, you can move it.

To Move a Category

- ▶ In the Categories pane, select the categories or descriptors that you would like to move into another category.
- ▶ From the menus, choose Categories > Move to Category. The menu presents a set of categories with the most recently created category at the top of the list. Select the name of the category to which you want to move the selected concepts.
 - If you see the name you are looking for, select it, and the selected elements are added to that category.
 - If you do not see it, select More to display the All Categories dialog box, and select the category from the list.

Figure 6-30
All Categories dialog box



Merging or Combining Categories

If you want to combine two or more existing categories into a new category, you can merge them. When you merge categories, a new category is created with a generic name. All of the concepts, types, and patterns used in the category descriptors are moved into this new category. You can later rename this category by editing the category properties. For more information, see the topic “Editing Category Properties” on p. 149.

To Merge a Category or Part of a Category

- ▶ In the Categories pane, select the elements you would like to merge together.
- ▶ From the menus, choose Categories > Merge Categories. The Category Properties dialog box is displayed in which you enter a name for the newly created category. The selected categories are merged into the new category as subcategories.

Forcing Responses into Categories

Forcing responses into and out of categories enables you to override the category definitions created by the automatic category building techniques without changing the actual category definition. You may find that although the response contains terms that are used to define a particular category, the response itself should not be in that category. In this case, you can force the response out of that category without having to remove the terms from the category definition.

Forcing is used in special cases where a response fits (or does not fit) a category but for one reason or another (for example, it contains a particular term) is assigned (or not assigned) to that category. Most typically, this occurs when a respondent uses sarcasm in his or her response, such as “*The pizza was great. I am sure everyone loves burnt, cold pizza.*” Let’s suppose that you had a category called Pos: [`<Food>` + `<Positive>`] to capture positive opinion regarding the food that a restaurant serves, this response may be assigned to that category. In this case, you might want to force this response out of the category.

To Force Responses Into or Out of Categories

- ▶ From within the Data pane, select the response that you want to force into or out of a particular category.
- ▶ From the menus choose Categories > Force Response Into or Categories > Force Response Out Of. A submenu displays the list of categories from which you can select.
- ▶ Select the category to which or from which you want to force this response. If you have created many categories, some may not be visible in the submenu.
 - In this case, choose More at the bottom of the submenu. The All Categories dialog box opens, in which you can select the category and click OK to apply the change.
 - If you want to force the response into a new category, select New Empty Category. A new category appears in the category tree using a generic name.

Whenever a category contains one or more forced responses a pseudo-category called either Force In or Force Out is displayed below the category name in the tree. You can also tell which responses are forced into or out of a category by displaying the Force In and Force Out columns in the Data pane. For more information, see the topic “The Data Pane” on p. 95.

To Clear a Forced Response State

- ▶ From within the Data pane, select the response that you no longer want to force into or out of a category.

- From the menus choose Categories > Force Response Into to force in, or choose Categories > Force Response Out Of to force out. The categories in which the response is forced out of or into are preceded by a check mark.

Figure 6-31
Forcing responses from within the Data pane

	Id	Response	Categories
1	1	little, light	
2	4	Having all my CDs in the palm of my hand!	
3	9	Small, great sound, capacity.	
4	16	It's fun to use	
5	18		
6	26	music with my friends and download internet.	friends
7	28	space for all of my CDs.	
8	31		battery life
9	33		
10	40		
11	45		
12	49		
13	53		
14	55		
15	63	it keeps playing. i do	
16	68	Large storage capa	

Menu Item	Submenu Item
Add to Category	
Add to Rule	Ctrl+U
Add to Annotation	
Force Response Into	
Force Response Out Of	
Mark Responses With	
Sort	
Select All	Ctrl+A
Copy	Ctrl+C
Find	Ctrl+F
Display Columns	
	New Empty Category...
	<input type="checkbox"/> design
	<input type="checkbox"/> songs
	<input type="checkbox"/> quality
	<input checked="" type="checkbox"/> feature
	<input type="checkbox"/> stereo
	<input type="checkbox"/> capacity
	<input type="checkbox"/> listening
	<input type="checkbox"/> tunes
	<input type="checkbox"/> music

- Select the category in the submenu that is checked and for which you want to remove the force. The check mark is removed and the response is no longer forced.

To Clear a Forced Response State

To clear all forced response states:

- From the menus choose Categories > Clear All > Force Ins or Categories > Clear All > Force Outs. The forced state on the responses is cleared and they are no longer forced into or out of the categories.

Text Matching in Categories

If you have tried forcing the extraction of a word or phrase through the linguistic resources and yet it is still not extracted due to other linguistic handling rules, you can create a text match entry to directly assign any categories containing that text into a particular category without using any extraction results.

When you add text match entries (a word or phrase) into a category, IBM® SPSS® Text Analytics for Surveys will automatically assign any responses containing the word or phrase to this category. Text matching should be used only if you have already tried to add this word to the linguistic resources in order to benefit most from this definition. For more information, see the topic “Forcing Terms” in Chapter 10 on p. 214.

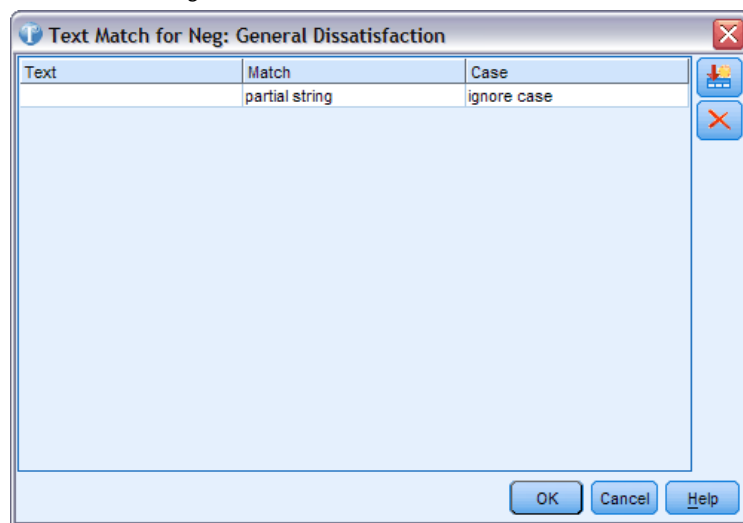
Whenever a text match is added to a category definition, a pseudo-category called Text Match is displayed below the category in the category tree. For more information, see the topic “The Data Pane” on p. 95.

You can also delete entries from this table by selecting the row(s) that you want to delete and clicking the Delete button.

To add a text match entry to a category definition:

- ▶ From within the Data pane, identify the word or phrase that you want to force into a category definition.
- ▶ In the Categories pane, select the category into which you want to force this word or phrase.
- ▶ From the menus choose Categories > Category Properties. The Category Properties dialog box opens.
- ▶ Click Advanced. The Text Match dialog box opens.

Figure 6-32
Text Match dialog box



- ▶ In the table, enter the word or phrase in the first cell in the Text column.
- ▶ Select how this word or phrase should be matched to text found in the responses. To match the word or phrase exactly as you have entered it, select the entire word or phrase. To match the word or phrase to longer phrases, select partial string.
- ▶ If the word or phrase you are entering is case sensitive, select match case in the Case column.
- ▶ Click OK to save your changes and to close the dialog box. The number of responses assigned to the category using text match entries is updated and displayed in the dialog box.
- ▶ Click OK to apply your changes.

Copying Categories

When you use the same or similar questions on one or more surveys, reusing the category definitions is a great time-saving option. You can copy the categories from one question to another in the same project. When you reuse categories, you will need to reextract in order to match the categories to the response data. Before you reextract, the categories will appear in the Categories pane with a question mark (?) for the frequency count.

Note: To reuse categories in another project, we recommend making a text analysis package with the categories and resources in your project and using this text analysis package (TAP) when creating your new project in the wizard. For more information, see the topic “Using Text Analysis Packages” in Chapter 3 on p. 40.

To Copy Categories within the Same Project

- ▶ Go to the question whose categories you want to copy.
- ▶ In the tree in the Categories pane, select all of the categories.
- ▶ From the menus, choose Edit > Copy to copy the categories.
- ▶ Go to the question (View > Question) into which you would like to paste these categories.
- ▶ From the menus, choose Edit > Paste to paste the categories. The categories are added to the pane. No frequency counts are known because you have not reextracted. Therefore, the counts appear as a question mark (?).

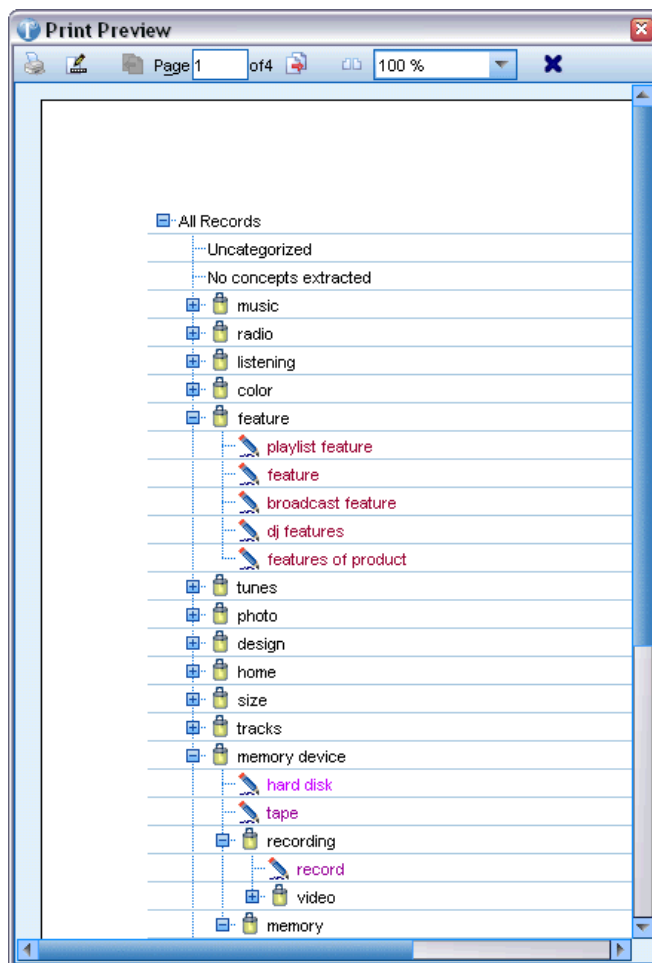
Printing Categories

You can print out the tree view in the Categories pane.

To Print the Category Tree View

- ▶ In the tree in the Categories pane, expand, collapse, or sort the tree elements according to what you want to see printed.
- ▶ From the menus choose File > Print > Print Categories. The Print Preview dialog box opens.

Figure 6-33
Print Preview dialog box



- ▶ Click the print button to print the view as it appears in the dialog box.

Deleting Categories

If you no longer want to keep a category, you can delete it. When you delete a category, any concepts that are not used in another category are visible on the Unused Extractions tab in the Extraction Results pane.

To Delete a Category

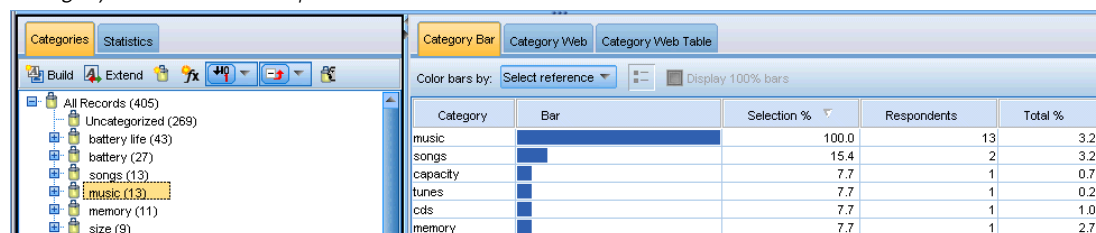
- ▶ In the Categories pane, select the category or categories that you would like to delete.
- ▶ From the menus, choose Edit > Delete.

Visualizing Graphs

When building your categories, it is important to take the time to review the category definitions, the responses they contain, and how the categories overlap. The visualization pane offers several perspectives on your categories. The Visualization pane is located in the upper right corner of the Question view. If it is not already visible, you can access this pane from the View menu (View > Panes > Visualization).

In this view, the visualization pane offers three perspectives on the commonalities in response categorization. The charts and graphs in this pane can be used to analyze your categorization results and aid in fine-tuning categories or reporting. When refining categories, you can use this pane to review your category definitions to uncover categories that are too similar (for example, they share more than 75% of their responses) or too distinct. If two categories are too similar, it might help you decide to combine the two categories. Alternatively, you might decide to refine the category definitions by removing certain descriptors from one category or the other. You can copy, paste, and print the results in this pane to help in your analysis or for reporting purposes.

Figure 7-1
Category and visualization panes



Depending on what is selected in the Extraction Results pane or Categories pane, you can view the corresponding interactions between responses and categories on each of the tabs in this pane. Each presents similar information but in a different manner or with a different level of detail. If necessary, you can customize the colors used in these graphs and charts in the Options dialog box. For more information, see the topic “Options: Display Tab” in Chapter 2 on p. 18.

Note: You can also generate summary graphs, such as a Top 5 Categories bar chart. These graphs, which are exported into HTML, can then be used in presentations. For more information, see the topic “Exporting Summary Graphs” in Chapter 4 on p. 58.

The Visualization pane offers the following graphs and charts:

- **Category Bar Chart.** A table and bar chart present the overlap between the responses corresponding to your selection and the associated categories. The bar chart also presents ratios of the responses in categories to the total number of responses. You can also select a reference variable, if you’ve imported any, to compare the reference variable values of the records in each category. For more information, see the topic “Category Bar Chart” on p. 160.

- **Category Web Graph.** This graph presents the response overlap for the categories to which the responses belong according to the selection in the other panes. For more information, see the topic “Category Web Graph” on p. 161.
- **Category Web Table.** This table presents the same information as the Category Web tab but in a table format. The table contains three columns that can be sorted by clicking the column headers. For more information, see the topic “Category Web Table” on p. 162.

For more information, see the topic “Categorizing Text Data” in Chapter 6 on p. 91.

Category Bar Chart

This tab displays a table and bar chart showing the overlap between the responses corresponding to your selection and the associated categories. The bar chart also presents ratios of the responses in categories to the total number of responses. You cannot edit the layout of this chart.

You can use the context menus in this bar chart to sort columns, change the graph colors, select the chart contents, copy the contents, as well as show or hide the legend.

The chart contains the following columns:

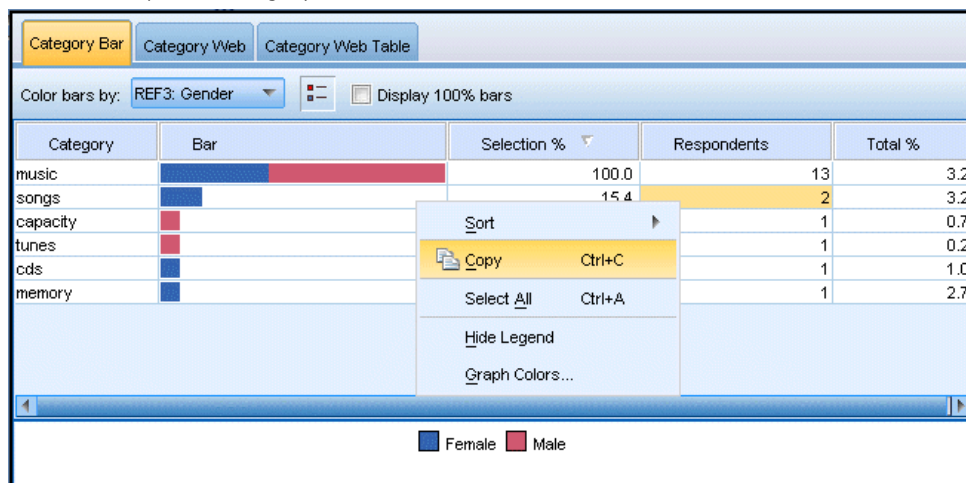
- **Category.** This column presents the name of the categories in your selection. By default, the most common category in your selection is listed first.
- **Bar.** This column presents, in a visual manner, the ratio of the records in a given category to the total number of records.
- **Selection %.** This column presents a percentage based on the ratio of the total number of records for a category to the total number of records represented in the selection.
- **Responses** This column presents the number of responses in a selection for the given category.
- **Total %.** This column presents a percentage based on the ratio of the total number of records for a given category compared to the total number of records for this question (not the selection).

You can also select an available reference variable from the dropdown list, to compare their values. When you select a reference variable, the bars in the table are divided and into color coded according to the values for the reference variables. By clicking on each colored reference value in a bar, the Data pane will update to show the subselection of responses according to the reference variable value. To see a legend for the reference variable values, click the Legend toolbar button.

Figure 7-2
Legend toolbar button



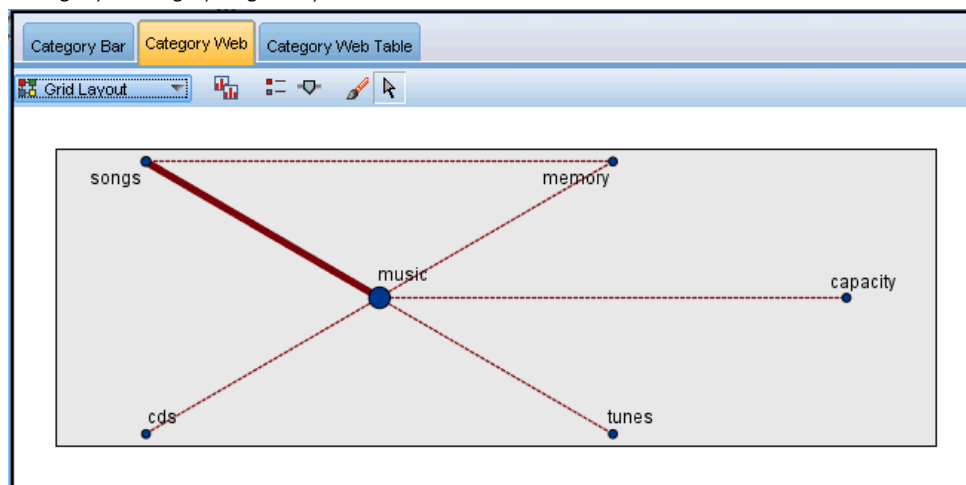
Figure 7-3
Visualization pane: Category bar



Category Web Graph

This tab displays a category web graph. The web presents the responses overlap for the categories to which the responses belong according to the selection in the other panes. If category labels exist, these labels appear in the graph. You can choose a graph layout (network, circle, directed, or grid) using the toolbar buttons in this pane.

Figure 7-4
Category Web graph, grid layout



In the web, each node represents a category. Using your mouse, you can select and move the nodes within the pane. The size of the node represents the relative size based on the number of records for that category in your selection. The thickness and color of the line between two categories denotes the number of common records they have. If you hover your mouse over a node in Explore mode, a ToolTip displays the name (or label) of the category and the overall number of records in the category.

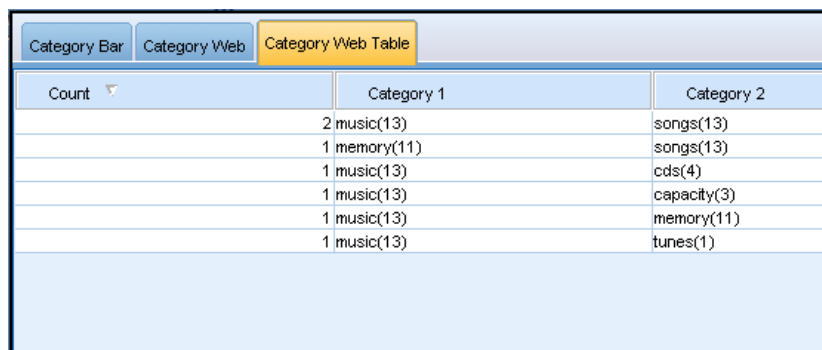
Note: By default, the Explore mode is enabled for the graphs on which you can move nodes. However, you can switch to Edit mode to edit your graph layouts including colors, fonts, legends, and more. For more information, see the topic “Using Graph Toolbars and Palettes” on p. 162.

Category Web Table

This tab displays the same information as the Category Web tab but in a table format. The table contains three columns that can be sorted by clicking the column headers:

- **Count.** This column presents the number of shared, or common, records between the two categories.
- **Category 1.** This column presents the name of the first category followed by the total number of records it contains, shown in parentheses.
- **Category 2.** This column presents the name of the second category followed by the total number of records it contains, shown in parentheses.

Figure 7-5
Visualization pane: Category web table



Count	Category 1	Category 2
2	music(13)	songs(13)
1	memory(11)	songs(13)
1	music(13)	cds(4)
1	music(13)	capacity(3)
1	music(13)	memory(11)
1	music(13)	tunes(1)






Using Graph Toolbars and Palettes

The category web graph has a toolbar that provides you with quick access to some common palettes from which you can perform a number of actions with your graphs. You can choose between the *Explore* view mode or the *Edit* view mode.

While Explore mode allows you to analytically explore the data and values represented by the visualization, Edit mode allows you to change the visualization’s layout and look. For example, you can change the fonts and colors to match your organization’s style guide. To select this mode, choose View > Visualization Pane > Edit Mode from the menus (or click the toolbar icon).

In Edit mode, there are several toolbars that affect different aspects of the visualization’s layout. If you find that there are any you don’t use, you can hide them to increase the amount of space in the dialog box in which the graph is displayed. To select or deselect toolbars, click on the relevant toolbar or palette name on the View menu.

Table 7-1
Text Analytics Toolbar buttons

Button/List	Description
	Enables Edit mode. Switch to the Edit mode to change the look of the graph, such as enlarging the font, changing the colors to match your corporate style guide, or removing labels and legends. For more information, see the topic “Editing Visualizations” on p. 163.
	Enables Explore mode. By default, the Explore mode is turned on, which means that you can move and drag nodes around the graph as well as hover over graph objects to reveal additional ToolTip information.
	Select a type of web display for the graphs. <ul style="list-style-type: none"> ■ Circle Layout. A general layout that can be applied to any graph. It lays out a graph assuming that links are undirected and treats all nodes the same. Nodes are only placed around the perimeter of a circle. ■ Network Layout. A general layout that can be applied to any graph. It lays out a graph assuming that links are undirected and treats all nodes the same. Nodes are placed freely within the layout. ■ Directed Layout. A layout that should only be used for directed graphs. This layout produces treelike structures from root nodes down to leaf nodes and organizes by colors. ■ Grid Layout. A general layout that can be applied to any graph. It lays out a graph assuming that links are undirected and treats all nodes the same. Nodes are only placed at grid points within the space.
	A toggle button that, when pressed, displays the legend. When the button is not pushed, the legend is not shown.
	A toggle button that, when pressed, displays the Links Slider beneath the graph. You can filter the results by sliding the arrow.

Editing Visualizations

You have several options for editing a visualization in **Edit mode**. You can:

- Edit text and format it.
- Change the fill color, transparency, and pattern of frames and graphic elements.
- Change the color and dashing of borders and lines.
- Rotate and change the shape and aspect ratio of point elements.
- Change the size of graphic elements (such as bars and points).
- Adjust the space around items by using margins and padding.
- Specify formatting for numbers.
- Change the axis and scale settings.
- Sort, exclude, and collapse categories on a categorical axis.
- Set the orientation of panels.
- Apply transformations to a coordinate system.
- Change statistics, graphic element types, and collision modifiers.
- Change the position of the legend.
- Apply visualization stylesheets.

The following topics describe how to perform these various tasks. It is also recommended that you read the general rules for editing graphs.

How to Switch to Edit Mode

- ▶ From the menus choose:
View > Edit Mode

General Rules for Editing Visualizations

Edit Mode

All edits are done in Edit mode. To enable Edit mode, from the menus choose:
View > Edit Mode

Selection

The options available for editing depend on selection. Different toolbar and properties palette options are enabled depending on what is selected. Only the enabled items apply to the current selection. For example, if an axis is selected, the Scale, Major Ticks, and Minor Ticks tabs are available in the properties palette.

Here are some tips for selecting items in the visualization:

- Click an item to select it.
- Select a graphic element (such as points in a scatterplot or bars in a bar chart) with a single click. After initial selection, click again to narrow the selection to groups of graphic elements or a single graphic element.
- Press Esc to deselect everything.

Palettes

When an item is selected in the visualization, the various palettes are updated to reflect the selection. The palettes contain controls for making edits to the selection. Palettes may be toolbars or a panel with multiple controls and tabs. Palettes can be hidden, so ensure the necessary palette is displayed for making edits. Check the View menu for palettes that are currently displayed.

You can reposition the palettes by clicking and dragging the empty space in a toolbar palette or the left side of other palettes. Visual feedback lets you know where you can dock the palette. For non-toolbar palettes, you can also click the close button to hide the palette and the undock button to display the palette in a separate window. Click the help button to display help for the specific palette.

Automatic Settings

Some settings provide an -auto- option. This indicates that automatic values are applied. Which automatic settings are used depends on the specific visualization and data values. You can enter a value to override the automatic setting. If you want to restore the automatic setting, delete the current value and press Enter. The setting will display -auto- again.

Removing/Hiding Items

You can remove/hide various items in the visualization. For example, you can hide the legend or axis label. To delete an item, select it and press Delete. If the item does not allow deletion, nothing will happen. If you accidentally delete an item, press Ctrl+Z to undo the deletion.

State

Some toolbars reflect the state of the current selection, others don't. The properties palette always reflects state. If a toolbar does *not* reflect state, this is mentioned in the topic that describes the toolbar.

Editing and Formatting Text

You can edit text in place and change the formatting of an entire text block. Note that you can't edit text that is linked directly to data values. For example, you can't edit a tick label because the content of the label is derived from the underlying data. However, you can format any text in the visualization.

How to Edit Text in Place

- ▶ Double-click the text block. This action selects all the text. All toolbars are disabled at this time, because you cannot change any other part of the visualization while editing text.
- ▶ Type to replace the existing text. You can also click the text again to display a cursor. Position the cursor in the desired position and enter the additional text.

How to Format Text

- ▶ Select the frame containing the text. Do not double-click the text.
- ▶ Format text using the font toolbar. If the toolbar is not enabled, make sure only the *frame* containing the text is selected. If the text itself is selected, the toolbar will be disabled.

Figure 7-6
Font toolbar



You can change the font:

- Color
- Family (for example, Arial or Verdana)
- Size (the unit is pt unless you indicate a different unit, such as pc)
- Weight
- Alignment relative to the text frame

Formatting applies to all the text in a frame. You can't change the formatting of individual letters or words in any particular block of text.

Changing Colors, Patterns, Dashings, and Transparency

Many different items in a visualization have a fill and border. The most obvious example is a bar in a bar chart. The color of the bars is the fill color. They may also have a solid, black border around them.

There are other less obvious items in the visualization that have fill colors. If the fill color is transparent, you may not know there is a fill. For example, consider the text in an axis label. It appears as if this text is “floating” text, but it actually appears in a frame that has a transparent fill color. You can see the frame by selecting the axis label.

Any frame in the visualization can have a fill and border style, including the frame around the whole visualization. Also, any fill has an associated opacity/transparency level that can be adjusted.

How to Change the Colors, Patterns, Dashing, and Transparency

- ▶ Select the item you want to format. For example, select the bars in a bar chart or a frame containing text. If the visualization is split by a categorical variable or field, you can also select the group that corresponds to an individual category. This allows you to change the default aesthetic assigned to that group. For example, you can change the color of one of the stacking groups in a stacked bar chart.
- ▶ To change the fill color, the border color, or the fill pattern, use the color toolbar.

Figure 7-7
Color toolbar

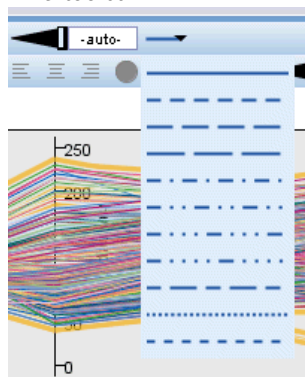


Note: This toolbar does not reflect the state of the current selection.

To change a color or fill, you can click the button to select the displayed option or click the drop-down arrow to choose another option. For colors, notice there is one color that looks like white with a red, diagonal line through it. This is the transparent color. You could use this, for example, to hide the borders on bars in a histogram.

- The first button controls the fill color.
 - The second button controls the border color.
 - The third button controls the fill pattern. The fill pattern uses the border color. Therefore, the fill pattern is visible only if there is a visible border color.
 - The fourth control is a slider and text box that control the opacity of the fill color and pattern. A lower percentage means less opacity and more transparency. 100% is fully opaque (no transparency).
- ▶ To change the dashing of a border or line, use the line toolbar.

Figure 7-8
Line toolbar



Note: This toolbar does not reflect the state of the current selection.

As with the other toolbar, you can click the button to select the displayed option or click the drop-down arrow to choose another option.

Rotating and Changing the Shape and Aspect Ratio of Point Elements

You can rotate point elements, assign a different predefined shape, or change the aspect ratio (the ratio of width to height).

How to Modify Point Elements

- ▶ Select the point elements. You cannot rotate or change the shape and aspect ratio of individual point elements.
- ▶ Use the symbol toolbar to modify the points.

Figure 7-9
Symbol toolbar



- The first button allows you to change the shape of the points. Click the drop-down arrow and select a predefined shape.
- The second button allows you to rotate the points to a specific compass position. Click the drop-down arrow and then drag the needle to the desired position.
- The third button allows you to change the aspect ratio. Click the drop-down arrow and then click and drag the rectangle that appears. The shape of the rectangle represents the aspect ratio.

Changing the Size of Graphic Elements

You can change the size of the graphic elements in the visualization. These include bars, lines, and points among others. If the graphic element is sized by a variable or field, the specified size is the *minimum* size.

How to Change the Size of the Graphic Elements

- ▶ Select the graphic elements you want to resize.
- ▶ Use the slider or enter a specific size for the option available on the symbol toolbar. The unit is pixels unless you indicate a different unit (see below for a full list of unit abbreviations). You can also specify a percentage (such as 30%), which means that a graphic element uses the specified percentage of the available space. The available space depends on the graphic element type and the specific visualization.

Table 7-2
Valid unit abbreviations

Abbreviation	Unit
cm	centimeter
in	inch
mm	millimeter
pc	pica
pt	point
px	pixel

Figure 7-10
Size control on symbol toolbar



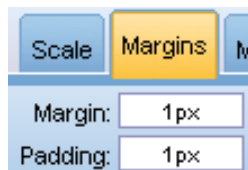
Specifying Margins and Padding

If there is too much or too little spacing around or inside a frame in the visualization, you can change its margin and padding settings. The **margin** is the amount of space between the frame and other items around it. The **padding** is the amount of space between the border of the frame and the *contents* of the frame.

How to Specify Margins and Padding

- ▶ Select the frame for which you want to specify margins and padding. This can be a text frame, the frame around the legend, or even the data frame displaying the graphic elements (such as bars and points).
- ▶ Use the Margins tab on the properties palette to specify the settings. All sizes are in pixels unless you indicate a different unit (such as cm or in).

Figure 7-11
Margins tab



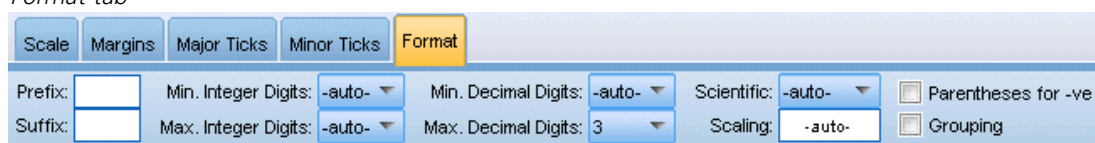
Formatting Numbers

You can specify the format for numbers in tick labels on a continuous axis or data value labels displaying a number. For example, you can specify that numbers displayed in the tick labels are shown in thousands.

How to Specify Number Formats

- ▶ Select the continuous axis tick labels or the data value labels if they contain numbers.
- ▶ Click the Format tab on the properties palette.

Figure 7-12
Format tab



- ▶ Select the desired number formatting options:

Prefix. A character to display at the beginning of the number. For example, enter a dollar sign (\$) if the numbers are salaries in U.S. dollars.

Suffix. A character to display at the end of the number. For example, enter a percentage sign (%) if the numbers are percentages.

Min. Integer Digits. Minimum number of digits to display in the integer part of a decimal representation. If the actual value does not contain the minimum number of digits, the integer part of the value will be padded with zeros.

Max. Integer Digits. Maximum number of digits to display in the integer part of a decimal representation. If the actual value exceeds the minimum number of digits, the integer part of the value will be replaced with asterisks.

Min. Decimal Digits. Minimum number of digits to display in the decimal part of a decimal or scientific representation. If the actual value does not contain the minimum number of digits, the decimal part of the value will be padded with zeros.

Max. Decimal Digits. Maximum number of digits to display in the decimal part of a decimal or scientific representation. If the actual value exceeds the minimum number of digits, the decimal is rounded to the appropriate number of digits.

Scientific. Whether to display numbers in scientific notation. Scientific notation is useful for very large or very small numbers. -auto- lets the application determine when scientific notation is appropriate.

Scaling. A scale factor, which is a number by which the original value is divided. Use a scale factor when the numbers are large, but you don't want the label to extend too much to accommodate the number. If you change the number format of the tick labels, be sure to edit the axis title to indicate how the number should be interpreted. For example, assume your scale axis displays salaries and the labels are 30,000, 50,000, and 70,000. You might enter a scale factor of 1000 to display 30, 50, and 70. You should then edit the scale axis title to include the text in thousands.

Parentheses for -ve. Whether parentheses should be displayed around negative values.

Grouping. Whether to display a character between groups of digits. Your computer's current locale determines which character is used for digit grouping.

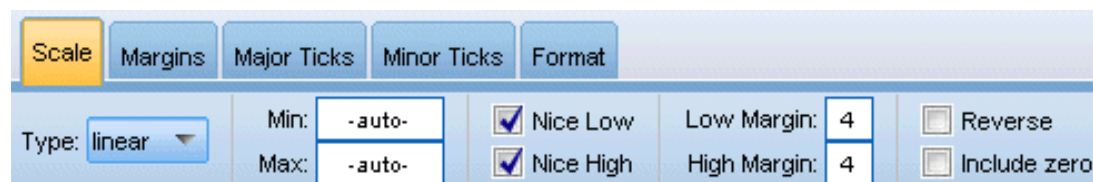
Changing the Axis and Scale Settings

There are several options for modifying axes and scales.

How to Change Axis and Scale Settings

- ▶ Select any part of the axis (for example, the axis label or tick labels).
- ▶ Use the Scale, Major Ticks, and Minor Ticks tabs on the properties palette to change the axis and scale settings.

Figure 7-13
Properties palette



Scale tab

Type. Specifies whether the scale is linear or transformed. Scale transformations help you understand the data or make assumptions necessary for statistical inference. On scatterplots, you might use a transformed scale if the relationship between the independent and dependent variables or fields is nonlinear. Scale transformations can also be used to make a skewed histogram more symmetric so that it resembles a normal distribution. Note that you are transforming only the scale on which the data are displayed; you are not transforming the actual data.

- **linear.** Specifies a linear, untransformed scale.
- **log.** Specifies a base-10 log transformed scale. To accommodate zero and negative values, this transformation uses a modified version of the log function. This “safe log” function is defined as $\text{sign}(x) * \log(1 + \text{abs}(x))$. So $\text{safeLog}(-99)$ equals:

$$\text{sign}(-99) * \log(1 + \text{abs}(-99)) = -1 * \log(1 + 99) = -1 * 2 = -2$$
- **power.** Specifies a power transformed scale, using an exponent of 0.5. To accommodate negative values, this transformation uses a modified version of the power function. This “safe power” function is defined as $\text{sign}(x) * \text{pow}(\text{abs}(x), 0.5)$. So $\text{safePower}(-100)$ equals:

$$\text{sign}(-100) * \text{pow}(\text{abs}(-100), 0.5) = -1 * \text{pow}(100, 0.5) = -1 * 10 = -10$$

Min/Max/Nice Low/Nice High. Specifies the range for the scale. Selecting Nice Low and Nice High allows the application to select an appropriate scale based on the data. The minimum and maximum are “nice” because they are typically whole values greater or less than the maximum and minimum data values. For example, if the data range from 4 to 92, a nice low and high for scale may be 0 and 100 rather than the actual data minimum and maximum. Be careful that you

don't set a range that is too small and hides important items. Also note that you cannot set an explicit minimum and maximum if the Include zero option is selected.

Low Margin/High Margin. Create margins at the low and/or high end of the axis. The margin appears perpendicular to the selected axis. The unit is pixels unless you indicate a different unit (such as cm or in). For example, if you set the High Margin to 5 for the vertical axis, a horizontal margin of 5 px runs along the top of the data frame.

Reverse. Specifies whether the scale is reversed.

Include zero. Indicates that the scale should include 0. This option is commonly used for bar charts to ensure the bars begin at 0, rather than a value near the height of the smallest bar. If this option is selected, Min and Max are disabled because you cannot set a custom minimum and maximum for the scale range.

Major Ticks/Minor Ticks Tabs

Ticks or **tick marks** are the lines that appear on an axis. These indicate values at specific intervals or categories. **Major ticks** are the tick marks with labels. These are also longer than other tick marks. **Minor ticks** are tick marks that appear between the major tick marks. Some options are specific to the tick type, but most options are available for major and minor ticks.

Show ticks. Specifies whether major or minor ticks are displayed on a graph.

Show gridlines. Specifies whether gridlines are displayed at the major or minor ticks. **Gridlines** are lines that cross a whole graph from axis to axis.

Position. Specifies the position of the tick marks relative to the axis.

Length. Specifies the length of the tick marks. The unit is pixels unless you indicate a different unit (such as cm or in).

Base. *Applies only to major ticks.* Specifies the value at which the first major tick appears.

Delta. *Applies only to major ticks.* Specifies the difference between major ticks. That is, major ticks will appear at every n th value, where n is the delta value.

Divisions. *Applies only to minor ticks.* Specifies the number of minor tick divisions between major ticks. The number of minor ticks is one less than the number of divisions. For example, assume that there are major ticks at 0 and 100. If you enter 2 as the number of minor tick divisions, there will be *one* minor tick at 50, dividing the 0–100 range and creating *two* divisions.

Editing Categories

You can edit the categories on a categorical axis in several ways:

- Change the sort order for displaying the categories.
- Exclude specific categories.
- Add a category that does not appear in the data set.
- Collapse/combine small categories into one category.

How to Change the Sort Order of Categories

- ▶ Select a categorical axis. The Categories palette displays the categories on the axis.

Note: If the palette is not visible, make sure that you have it enabled.

- ▶ In the Categories palette, select a sorting option from the drop-down list:

Custom. Sort categories based on the order in which they appear in the palette. Use the arrow buttons to move categories to the top of the list, up, down, and to the bottom of the list.

Data. Sort categories based on the order in which they occur in the dataset.

Name. Sort categories alphabetically, using the names as displayed in the palette.

Value. Sort categories by the underlying data value, using the values displayed in parentheses in the palette. Only data sources with metadata (such as IBM® SPSS® Statistics data files) support this option.

Statistic. Sort categories based on the calculated statistic for each category. Examples of statistics include counts, percentages, and means. This option is available only if a statistic is used in the graph.

How to Add a Category

By default, only categories that appear in the data set are available. You can add a category to the visualization if needed.

- ▶ Select a categorical axis. The Categories palette displays the categories on the axis.

Note: If the palette is not visible, make sure that you have it enabled.

- ▶ In the Categories palette, click the add category button:

Figure 7-14
Add category button



- ▶ In the Add a new category dialog box, enter a name for the category.
- ▶ Click OK.

How to Exclude Specific Categories

- ▶ Select a categorical axis. The Categories palette displays the categories on the axis.

Note: If the palette is not visible, make sure that you have it enabled.

- ▶ In the Categories palette, select a category name in the Include list, and then click the X button. To move the category back, select its name in the Excluded list, and then click the arrow to the right of the list.

How to Collapse/Combine Small Categories

You can combine categories that are so small you don't need to display them separately. For example, if you have a pie chart with many categories, consider collapsing categories with a percentage less than 10. Collapsing is available only for statistics that are additive. For example, you can't add means together because means are not additive. Therefore, combining/collapsing categories using a mean is not available.

- ▶ Select a categorical axis. The Categories palette displays the categories on the axis.
Note: If the palette is not visible, make sure that you have it enabled.
- ▶ In the Categories palette, select Collapse and specify a percentage. Any categories whose percentage of the total is less than the specified number are combined into one category. The percentage is based on the statistic shown in the chart. Collapsing is available only for count-based and summation (sum) statistics.

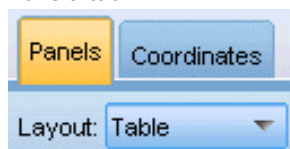
Changing the Orientation Panels

If you are using panels in your visualization, you can change their orientation.

How to Change the Orientation of the Panels

- ▶ Select any part of the visualization.
- ▶ Click the Panels tab on the properties palette.

Figure 7-15
Panels tab



- ▶ Select an option from Layout:
 - Table.** Lays out panels like a table, in that there is a row or column assigned to every individual value.
 - Transposed.** Lays out panels like a table, but also swaps the original rows and columns. This option is not the same as transposing the graph itself. Note that the x axis and the y axis are unchanged when you select this option.
 - List.** Lays out panels like a list, in that each cell represents a combination of values. Columns and rows are no longer assigned to individual values. This option allows the panels to wrap if needed.

Transforming the Coordinate System

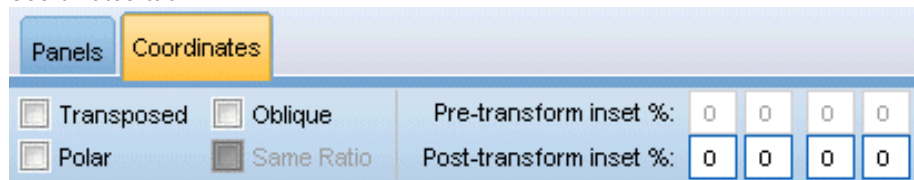
Many visualizations are displayed in a flat, rectangular coordinate system. You can transform the coordinate system as needed. For example, you can apply a polar transformation to the coordinate system, add oblique drop shadow effects, and transpose the axes. You can also undo any of these

transformations if they are already applied to the current visualization. For example, a pie chart is drawn in a polar coordinate system. If desired, you can undo the polar transformation and display the pie chart as a single stacked bar in a rectangular coordinate system.

How to Transform the Coordinate System

- ▶ Select the coordinate system that you want to transform. You select the coordinate system by selecting the frame around the individual graph.
- ▶ Click the Coordinates tab on the properties palette.

Figure 7-16
Coordinates tab



- ▶ Select the transformations that you want to apply to the coordinate system. You can also deselect a transformation to undo it.

Transposed. Changing the orientation of the axes is called **transposing**. It is similar to swapping the vertical and horizontal axes in a 2-D visualization.

Polar. A polar transformation draws the graphic elements at a specific angle and distance from the center of the graph. A pie chart is a 1-D visualization with a polar transformation that draws the individual bars a specific angles. A radar chart is a 2-D visualization with a polar transformation that draws graphic elements a specific angles and distances from the center of the graph. A 3-D visualization would also include an additional depth dimension.

Oblique. An oblique transformation adds a 3-D effect to the graphic elements. This transformation adds depth to the graphic elements, but the depth is purely decorative. It is not influenced by particular data values.

Same Ratio. Applying the same ratio specifies that the same distance on each scale represents the same difference in data values. For example, 2cm on both scales represent a difference of 1000.

Pre-transform inset %. If axes are clipped after the transformation, you may want to add insets to the graph before applying the transformation. The insets shrink the dimensions by a certain percentage before any transformations are applied to the coordinate system. You have control over the lower x , upper x , lower y , and upper y dimensions, in that order.

Post-transform inset %. If you want to change the aspect ratio of the a graph, you can add insets to the graph after applying the transformation. The insets shrink the dimensions by a certain percentage after any transformations are applied to the coordinate system. These insets can also be applied even if no transformation is applied to the graph. You have control over the lower x , upper x , lower y , and upper y dimensions, in that order.

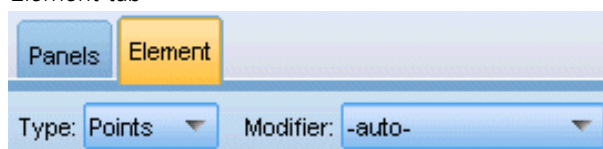
Changing Statistics and Graphic Elements

You can convert a to another type, change the statistic used to draw the graphic element, or specify the collision modifier that determines what happens when graphic elements overlap.

How to Convert a Graphic Element

- ▶ Select the graphic element that you want to convert.
- ▶ Click the Element tab on the properties palette.

Figure 7-17
Element tab



- ▶ Select a new graphic element type from the Type list.

Graphic Element Type	Description
Point	A marker identifying a specific data point. A point element is used in scatterplots and other related visualizations.
Interval	A rectangular shape drawn at a specific data value and filling the space between an origin and another data value. An interval element is used in bar charts and histograms.
Line	A line that connects data values.
Path	A line that connects data values in the order they appear in the dataset.
Area	A line that connects data elements with the area between the line and an origin filled in.
Polygon	A multi-sided shape enclosing a data region. A polygon element could be used in a binned scatterplot or a map.
Schema	An element consisting of a box with whiskers and markers indicating outliers. A schema element is used for boxplots.

How to Change the Statistic

- ▶ Select the graphic element whose statistic you want to change.
- ▶ Click the Element tab on the properties palette.
- ▶ From the Summary drop-down list, select a new statistic. Note that selecting a statistic aggregates the data. If instead you want the visualization to display unaggregated data, select (no statistic) from the Summary list.

Summary Statistics Calculated from a Continuous Field

- **Mean.** A measure of central tendency. The arithmetic average, the sum divided by the number of cases.
- **Median.** The value above and below which half of the cases fall, the 50th percentile. If there is an even number of cases, the median is the average of the two middle cases when they are sorted in ascending or descending order. The median is a measure of central tendency not

sensitive to outlying values (unlike the mean, which can be affected by a few extremely high or low values).

- **Mode.** The most frequently occurring value. If several values share the greatest frequency of occurrence, each of them is a mode.
- **Minimum.** The smallest value of a numeric variable.
- **Maximum.** The largest value of a numeric variable.
- **Range.** The difference between the minimum and maximum values.
- **Mid Range.** The middle of the range, that is, the value whose difference from the minimum is equal to its difference from the maximum.
- **Sum.** The sum or total of the values, across all cases with nonmissing values.
- **Cumulative Sum.** The cumulative sum of the values. Each graphic element shows the sum for one subgroup plus the total sum of all previous groups.
- **Percent Sum.** The percentage within each subgroup based on a summed field compared to the sum across all groups.
- **Cumulative Percent Sum.** The cumulative percentage within each subgroup based on a summed field compared to the sum across all groups. Each graphic element shows the percentage for one subgroup plus the total percentage of all previous groups.
- **Variance.** A measure of dispersion around the mean, equal to the sum of squared deviations from the mean divided by one less than the number of cases. The variance is measured in units that are the square of those of the variable itself.
- **Standard Deviation.** A measure of dispersion around the mean. In a normal distribution, 68% of cases fall within one standard deviation of the mean and 95% of cases fall within two standard deviations. For example, if the mean age is 45, with a standard deviation of 10, 95% of the cases would be between 25 and 65 in a normal distribution.
- **Standard Error.** A measure of how much the value of a test statistic varies from sample to sample. It is the standard deviation of the sampling distribution for a statistic. For example, the standard error of the mean is the standard deviation of the sample means.
- **Kurtosis.** A measure of the extent to which observations cluster around a central point. For a normal distribution, the value of the kurtosis statistic is zero. Positive kurtosis indicates that, relative to a normal distribution, the observations are more clustered about the center of the distribution and have thinner tails until the extreme values of the distribution, at which point the tails of the leptokurtic distribution are thicker relative to a normal distribution. Negative kurtosis indicates that, relative to a normal distribution, the observations cluster less and have thicker tails until the extreme values of the distribution, at which point the tails of the platykurtic distribution are thinner relative to a normal distribution.
- **Skewness.** A measure of the asymmetry of a distribution. The normal distribution is symmetric and has a skewness value of 0. A distribution with a significant positive skewness has a long right tail. A distribution with a significant negative skewness has a long left tail. As a guideline, a skewness value more than twice its standard error is taken to indicate a departure from symmetry.

The following region statistics may result in more than one graphic element per subgroup. When using the interval, area, or edge graphic elements, a region statistic results in one graphic element showing the range. All other graphic elements result in two separate elements, one showing the start of the range and one showing the end of the range.

- **Region: Range.** The range of values between the minimum and maximum values.
- **Region: 95% Confidence Interval of Mean.** A range of values that has a 95% chance of including the population mean.
- **Region: 95% Confidence Interval of Individual.** A range of values that has a 95% chance of including the predicted value given the individual case.
- **Region: 1 Standard Deviation above/below Mean.** A range of values between 1 **standard deviation** above and below the **mean**.
- **Region: 1 Standard Error above/below Mean.** A range of values between 1 **standard error** above and below the **mean**.

Count-Based Summary Statistics

- **Count.** The number of rows/cases.
- **Cumulative Count.** The cumulative number of rows/cases. Each graphic element shows the count for one subgroup plus the total count of all previous groups.
- **Percent of Count.** The percentage of rows/cases in each subgroup compared to the total number of rows/cases.
- **Cumulative Percent of Count.** The cumulative percentage of rows/cases in each subgroup compared to the total number of rows/cases. Each graphic element shows the percentage for one subgroup plus the total percentage of all previous groups.

How to Specify the Collision Modifier

The collision modifier determines what happens when graphic elements overlap.

- ▶ Select the graphic element for which you want to specify the collision modifier.
- ▶ Click the Element tab on the properties palette.
- ▶ From the Modifier drop-down list, select a collision modifier. -auto- lets the application determine which collision modifier is appropriate for the graphic element type and statistic.

Overlay. Draw graphic elements on top of each other when they have the same value.

Stack. Stack graphic elements that would normally be superimposed when they have the same data values.

Dodge. Move graphic elements next to other graphic elements that appear at the same value, rather than superimposing them. The graphic elements are arranged symmetrically. That is, the graphic elements are moved to opposite sides of a central position. Dodging is very similar to clustering.

Pile. Move graphic elements next to other graphic elements that appear at the same value, rather than superimposing them. The graphic elements are arranged asymmetrically. That is, the graphic elements are piled on top of one another, with the graphic element on the bottom positioned at a specific value on the scale.

Jitter (normal). Randomly reposition graphic elements at the same data value using a normal distribution.

Jitter (uniform). Randomly reposition graphic elements at the same data value using a uniform distribution.

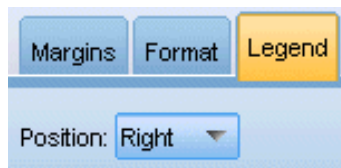
Changing the Position of the Legend

If a graph includes a legend, the legend is typically displayed to the right of a graph. You can change this position if needed.

How to Change the Legend Position

- ▶ Select the legend.
- ▶ Click the Legend tab on the properties palette.

Figure 7-18
Legend tab



- ▶ Select a position.

Copying a Visualization and Visualization Data

The General palette includes buttons for copying the visualization and its data.

Figure 7-19
Copy visualization button



Copying the visualization. This action copies the visualization to the clipboard as an image. Multiple image formats are available. When you paste the image into another application, you can choose a “paste special” option to select one of the available image formats for pasting.

Figure 7-20
Copy visualization data button



Copying the visualization data. This action copies the underlying data that is used to draw the visualization. The data is copied to the clipboard as plain text or HTML-formatted text. When you paste the data into another application, you can choose a “paste special” option to choose one of these formats for pasting.

Keyboard Shortcuts

Table 7-3
Keyboard shortcuts

Shortcut Key	Function
Ctrl+Space	Toggle between Explore and Edit mode
Delete	Delete a visualization item
Ctrl+Z	Undo
Ctrl+Y	Redo
F2	Display outline for selecting items in the graph

Part III:
Resource Editor

Templates and Resources

IBM® SPSS® Text Analytics for Surveys rapidly and accurately captures and extracts key concepts from text data. This extraction process relies heavily on linguistic resources to dictate how to extract information from text data. For more information, see the topic “How Extraction Works” in Chapter 1 on p. 3. You can fine-tune these resources in the Resource Editor view.

When you install the software, you also get a set of specialized resources. These shipped resources allow you to benefit from years of research and fine-tuning for specific languages and specific applications. Since the shipped resources may not always be perfectly adapted to the context of your data, you can edit these resource templates or even create and use custom libraries uniquely fine-tuned to your organization’s data. These resources come in various forms and each can be used in your project. Resources can be found in the following:

- **Resource templates.** Templates are made up of a set of libraries, types, and some advanced resources which together form a specialized set of resources adapted to a particular domain or context such as product opinions.
- **Text analysis packages (TAP).** In addition to the resources stored in a template, TAPs also bundle together one or more specialized category sets generated using those resources so that both the categories and the resources are stored together and reusable. For more information, see the topic “Using Text Analysis Packages” in Chapter 3 on p. 40.
- **Libraries.** Libraries are used as building blocks for both TAPs and templates. They can also be added individually to resources in your project. Each library is made up of several dictionaries used to define and manage types, synonyms, and exclude lists. While libraries are also delivered individually, they are prepackaged together in templates and TAPs. For more information, see the topic “Working with Libraries” in Chapter 9 on p. 195.

Note: During extraction, some compiled internal resources are also used. These compiled resources contain a large number of definitions complementing the types in the Core library. These compiled resources cannot be edited.

The Resource Editor offers access to the set of resources used to produce the extraction results (concepts, types, and patterns). There are a number of tasks you might perform in the Resource Editor and they include:

- **Working with libraries.** For more information, see the topic “Working with Libraries” in Chapter 9 on p. 195.
- **Creating type dictionaries.** For more information, see the topic “Creating Types” in Chapter 10 on p. 209.
- **Adding terms to dictionaries.** For more information, see the topic “Adding Terms” in Chapter 10 on p. 210.
- **Creating synonyms.** For more information, see the topic “Defining Synonyms” in Chapter 10 on p. 218.
- **Updating the resources in TAPs.** For more information, see the topic “Updating Text Analysis Packages” in Chapter 3 on p. 42.

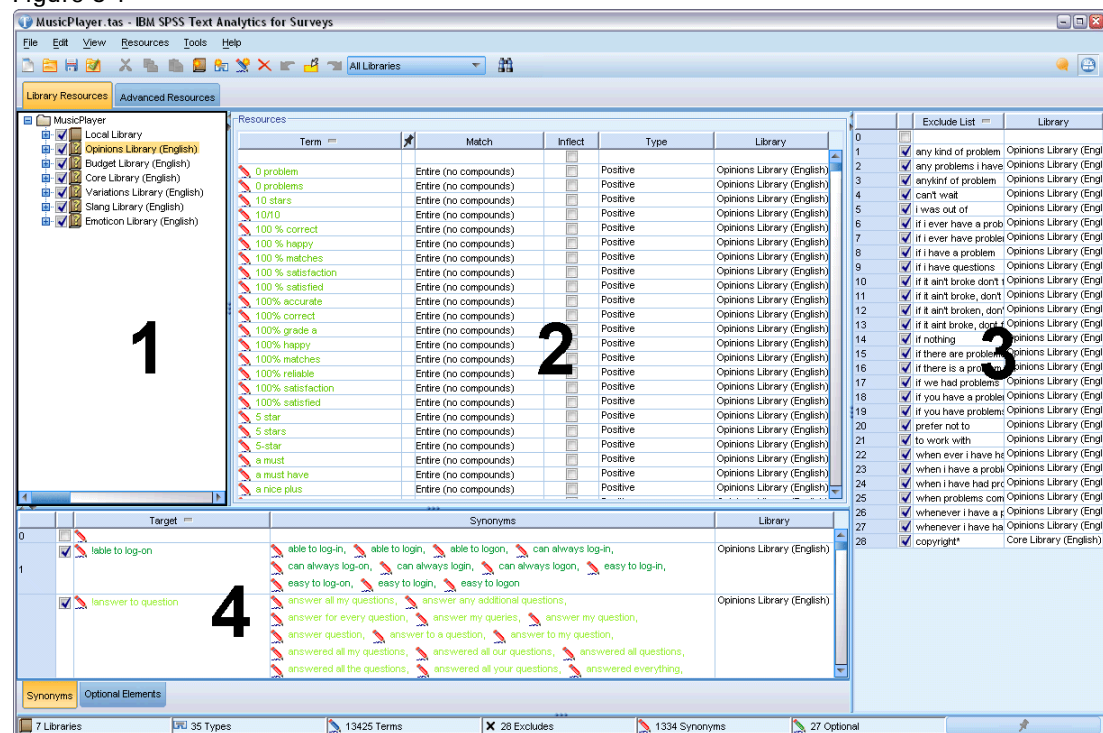
- **Making templates.** For more information, see the topic “Making and Updating Templates” on p. 186.
- **Importing and exporting templates.** For more information, see the topic “Importing and Exporting Templates” on p. 189.
- **Publishing libraries.** For more information, see the topic “Publishing Libraries” in Chapter 9 on p. 204.

The Editor Interface

The operations that you perform in the Resource Editor revolve around the management and fine-tuning of the linguistic resources. These resources are stored in the form of templates and libraries. For more information, see the topic “Type Dictionaries” in Chapter 10 on p. 207.

Library Resources tab

Figure 8-1



The interface is organized into four parts, as follows:

1. Library Tree pane. Located in the upper left corner, this pane displays a tree of the libraries. You can enable and disable libraries in this tree as well as filter the views in the other panes by selecting a library in the tree. You can perform many operations in this tree using the context menus. If you expand a library in the tree, you can see the set of types it contains. You can also filter this list through the View menu if you want to focus on a particular library only.

2. Term Lists from Type Dictionaries pane. Located to the right of the library tree, this pane displays the term lists of the type dictionaries for the libraries selected in the tree. A **type dictionary** is a collection of terms to be grouped under one label, or type, name. When the extraction engine reads your text data, it compares words found in the text to the terms in the type dictionaries. If an extracted concept appears as a term in a type dictionary, then that type name is assigned. You can think of the type dictionary as a distinct dictionary of terms that have something in common. For example, the <Location> type in the Core library contains concepts such as `new orleans`, `great britain`, and `new york`. These terms all represent geographical locations. A library can contain one or more type dictionaries. For more information, see the topic “Type Dictionaries” in Chapter 10 on p. 207.

3. Exclude Dictionary pane. Located on the right side, this pane displays the collection of terms that will be excluded from the final extraction results. The terms appearing in this exclude dictionary do not appear in the Extraction Results pane. Excluded terms can be stored in the library of your choosing. However, the Exclude Dictionary pane displays all of the excluded terms for all libraries visible in the library tree. For more information, see the topic “Exclude Dictionaries” in Chapter 10 on p. 222.

4. Substitution Dictionary pane. Located in the lower left, this pane displays synonyms and optional elements, each in their own tab. Synonyms and optional elements help group similar terms under one lead, or target, concept in the final extraction results. This dictionary can contain known synonyms and user-defined synonyms and elements, as well as common misspellings paired with the correct spelling. Synonym definitions and optional elements can be stored in the library of your choosing. However, the substitution dictionary pane displays all of the contents for all libraries visible in the library tree. While this pane displays all synonyms or optional elements from all libraries, The substitutions for all of the libraries in the tree are shown together in this pane. A library can contain only one substitution dictionary. For more information, see the topic “Substitution/Synonym Dictionaries” in Chapter 10 on p. 217.

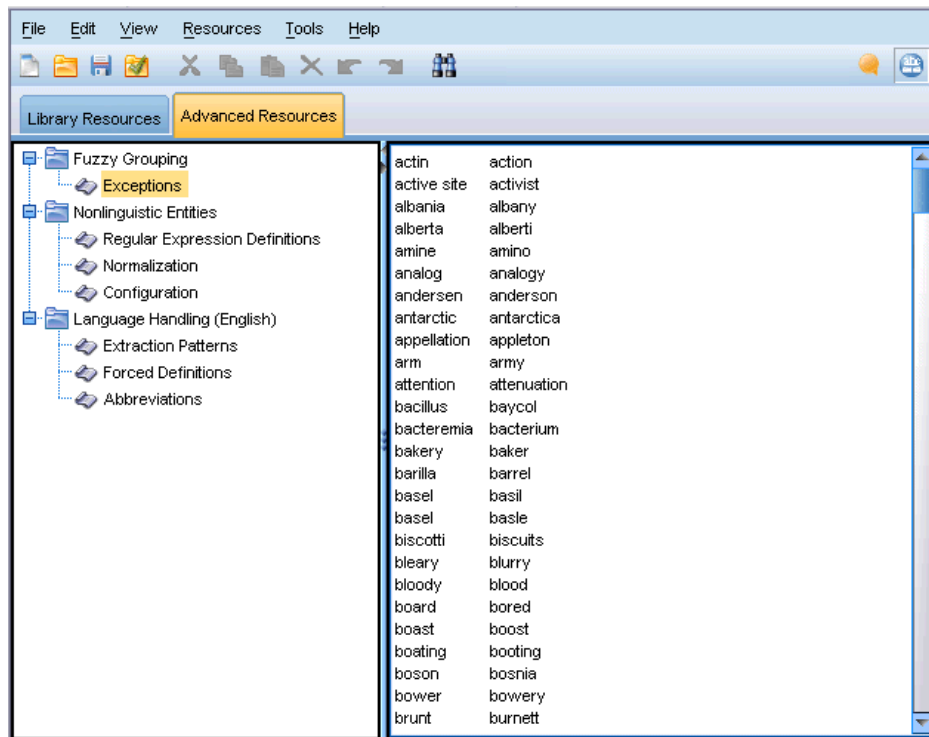
Note:

- If you want to filter so that you see only the information pertaining to a single library, you can change the library view using the drop-down list on the toolbar. It contains a top-level entry called All Libraries as well as an additional entry for each individual library. For more information, see the topic “Viewing Libraries” in Chapter 9 on p. 199.

Advanced Resources tab

The advanced resources are available from the second tab of the editor view. You can review and edit the advanced resources in this tab. For more information, see the topic “About Advanced Resources” in Chapter 11 on p. 225.

Figure 8-2
Advanced Resources

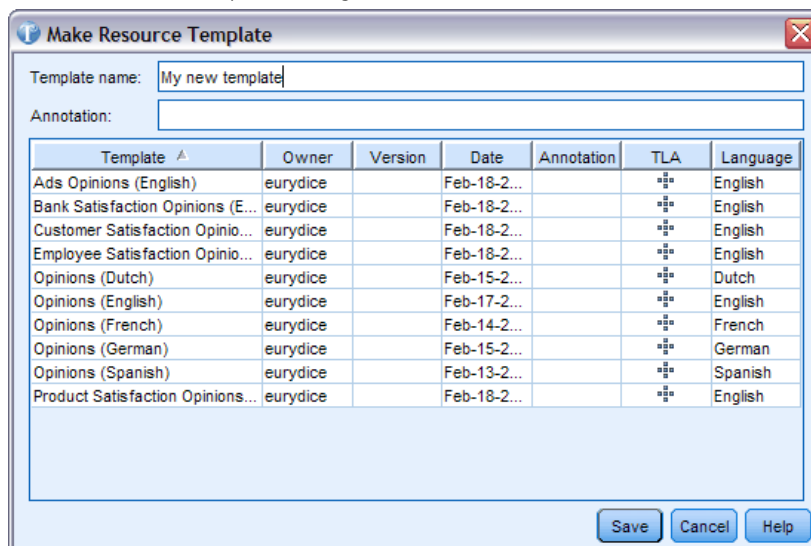


Making and Updating Templates

Whenever you make changes to your resources and want to reuse them in the future, you can save the resources as a template. When doing so, you can choose to save using an existing template name or by providing a new name. Then, whenever you load this template in the future, you'll be able to obtain the same resources.

Note: You can also publish and share your libraries. For more information, see the topic "Sharing Libraries" in Chapter 9 on p. 202.

Figure 8-3
Make Resource Template dialog box



To Make (or Update) a Template

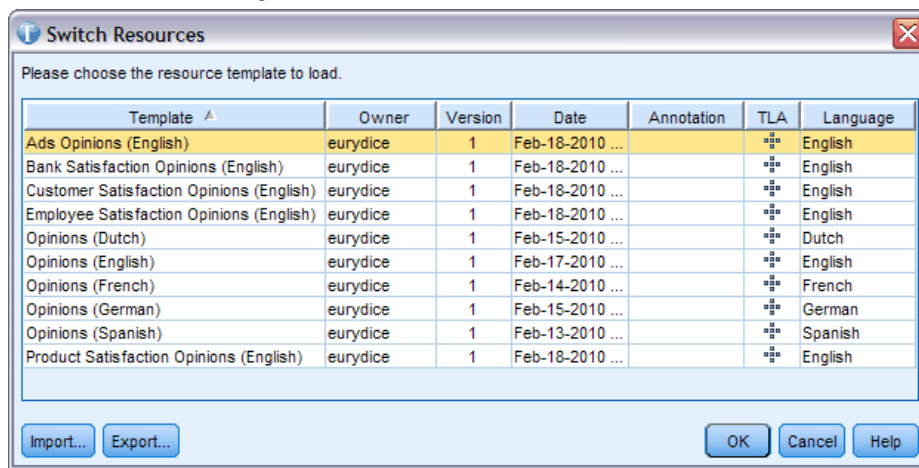
- ▶ From the menus in the Resource Editor view, choose Resources > Make Resource Template. The Make Resource Template dialog box opens.
- ▶ Enter a new name in the Template Name field, if you want to make a new template. Select a template in the table, if you want to overwrite an existing template with the currently loaded resources.
- ▶ Click Save to make the template.

Switching Resource Templates

If you want to replace the resources currently loaded with a copy of those from another template, you can switch to those resources. Doing so will overwrite any resources currently loaded.

You can select the template whose contents you want copy into the Resource Editor and click OK. This replaces the resources you have in this project.

Figure 8-4
Switch Resources dialog box



To Switch Resources

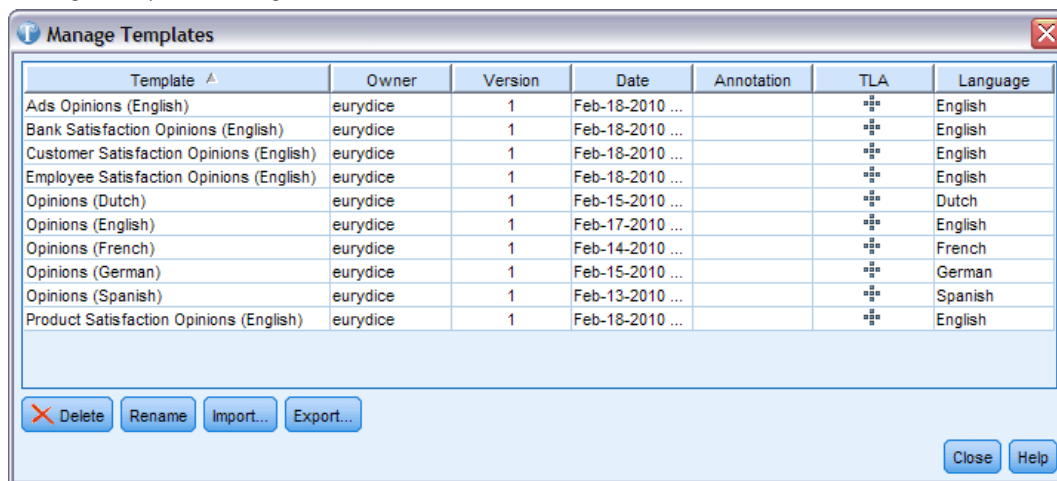
- ▶ From the menus in the Resource Editor view, choose Resources > Switch Resource Templates. The Switch Resource Templates dialog box opens.
- ▶ Select the template you want to use from those shown in the table.
- ▶ Click OK to abandon those resources currently loaded and load a copy of those in the selected template in their place. If you have made changes to your resources and want to save your libraries for a future use, you can publish, update, and share them before switching. For more information, see the topic “Sharing Libraries” in Chapter 9 on p. 202.

Managing Templates

There are also some basic management tasks you might want to perform from time to time on your templates, such as renaming your templates, importing and exporting templates, or deleting obsolete templates. These tasks are performed in the Manage Templates dialog box. Importing and exporting templates enables you to share templates with other users. For more information, see the topic “Importing and Exporting Templates” on p. 189.

Note: You cannot rename or delete the templates that are installed (or shipped) with this product. Instead, if you want to rename, you can open the installed template and make a new one with the name of your choice. You can delete your custom templates; however, if you try to delete a shipped template, it will be reset to the version originally installed.

Figure 8-5
Manage Templates dialog box



To Rename a Template

- ▶ From the menus, choose Resources > Manage Resource Templates. The Manage Templates dialog box opens.
- ▶ Select the template you want to rename and click Rename. The name box becomes an editable field in the table.
- ▶ Type a new name and press the Enter key. A confirmation dialog box opens.
- ▶ If you are satisfied with the name change, click Yes. If not, click No.

To Delete a Template

- ▶ From the menus, choose Resources > Manage Resource Templates. The Manage Templates dialog box opens.
- ▶ In the Manage Templates dialog box, select the template you want to delete.
- ▶ Click Delete. A confirmation dialog box opens.
- ▶ Click Yes to delete or click No to cancel the request. If you click Yes, the template is deleted.

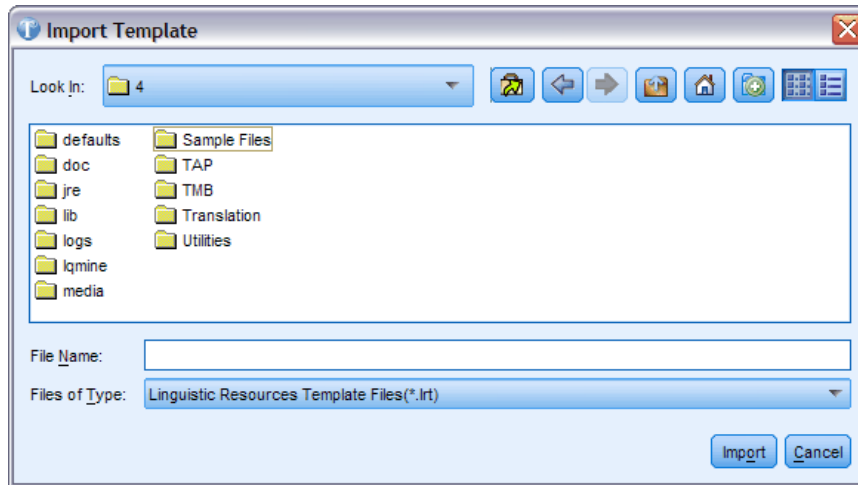
Importing and Exporting Templates

You can share templates with other users or machines by importing and exporting them. Templates are stored in an internal database but can be exported as *.lrt files to your hard drive. You can import and export templates in the Manage Templates dialog box in the Resource Editor.

To Import a Template

- ▶ In the dialog box, click Import. The Import Template dialog box opens.

Figure 8-6
Import Template dialog box

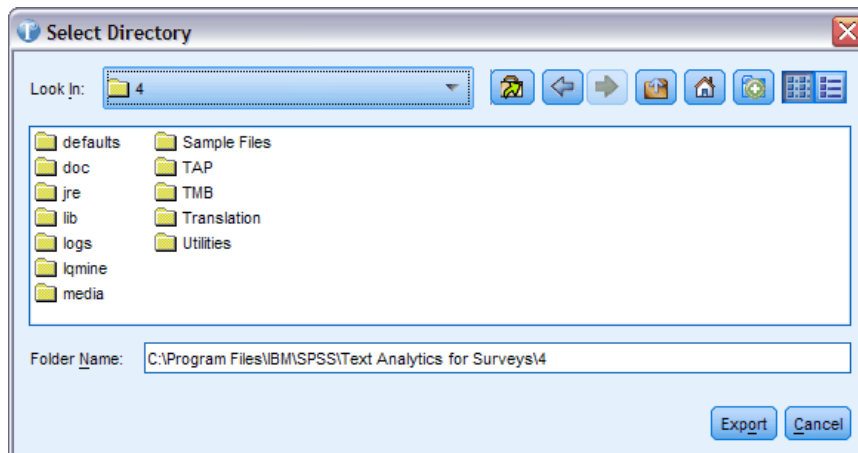


- ▶ Select the resource template file (*.lrt) to import and click Import. You can save the template you are importing with another name or overwrite the existing one. The dialog box closes, and the template now appears in the table.

To Export a Template

- ▶ In the dialog box, select the template you want export and click Export. The Select Directory dialog box opens.

Figure 8-7
Select Directory dialog box



- ▶ Select the directory to which you want to export and click Export. This dialog box closes, and the template is exported and carries the file extension (*.lrt)

Backing Up Resources

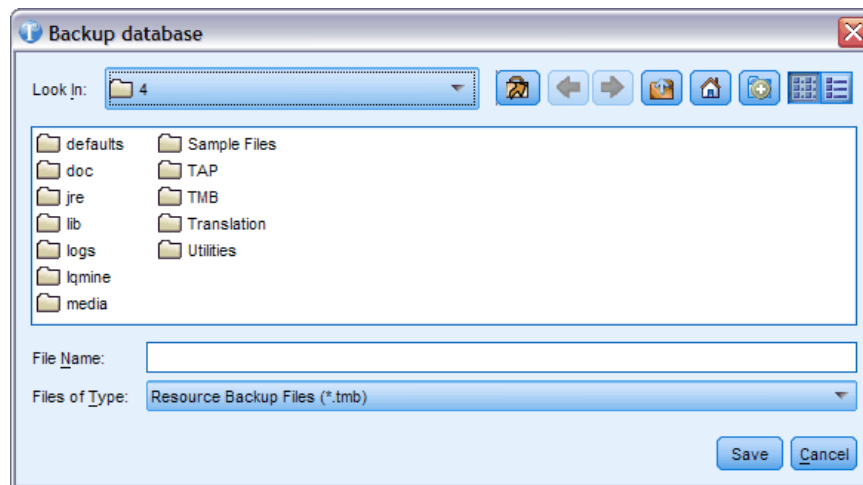
You may want to back up your resources from time to time as a security measure.

Important! When you restore, the entire contents of your resources will be wiped clean and only the contents of the backup file will be accessible in the product. This includes any open work.

To Back Up the Resources

- ▶ From the menus, choose Resources > Backup Tools > Backup Resources. The Backup dialog box opens.

Figure 8-8
Backup Resources dialog box

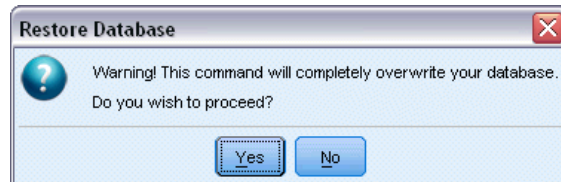


- ▶ Enter a name for your backup file and click Save. The dialog box closes, and the backup file is created.

To Restore the Resources

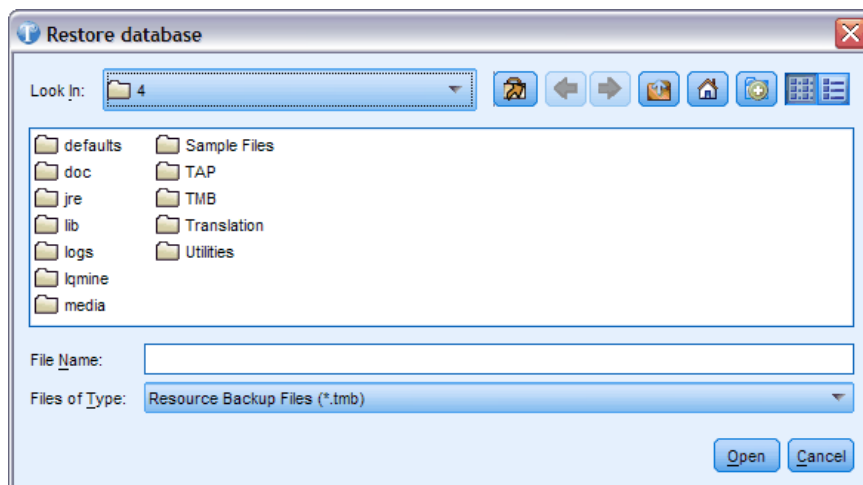
- ▶ From the menus, choose Resources > Backup Tools > Restore Resources. An alert warns you that restoring will overwrite the current contents of your database.

Figure 8-9
Overwrite warning message



- ▶ Click Yes to proceed. If you have a project open, it will be kept, since it is in memory; however, you must save it again to keep it in the newly restored database. The dialog box opens.

Figure 8-10
Restore Resources dialog box



- ▶ Select the backup file you want to restore and click Open. The dialog box closes, and resources are restored in the application.

Important! When you restore, the entire contents of your resources will be wiped clean and only the contents of the backup file will be accessible in the product. This includes any open work.

Importing Resource Files

If you have made changes directly in resource files outside of this product, you can import them into a selected library by selecting that library and proceeding with the import. When you import a directory, you can import all of supported files into a specific open library as well. You can only import **.txt* files.

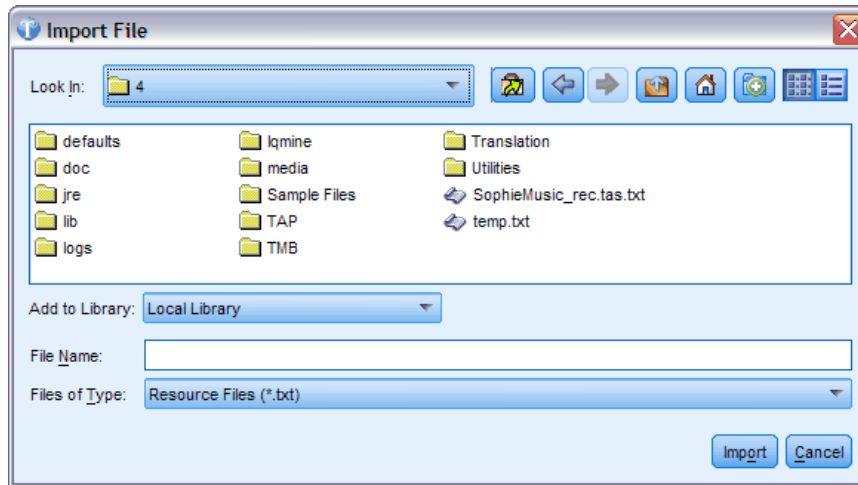
Each imported file must contain only one entry per line, and if the contents are structured as:

- A list words or phrases (one per line). The file is imported as a term list for a type dictionary, where the type dictionary takes the name of the file minus the extension.
- A list of entries such as *term1*<TAB>*term2*, then it is imported as a list of synonyms, where *term1* is the set of the underlying term and *term2* is the target term.

To Import a Single Resource File

- ▶ From the menus, choose Resources > Import Files > Import Single File. The Import File dialog box opens.

Figure 8-11
Import File dialog box

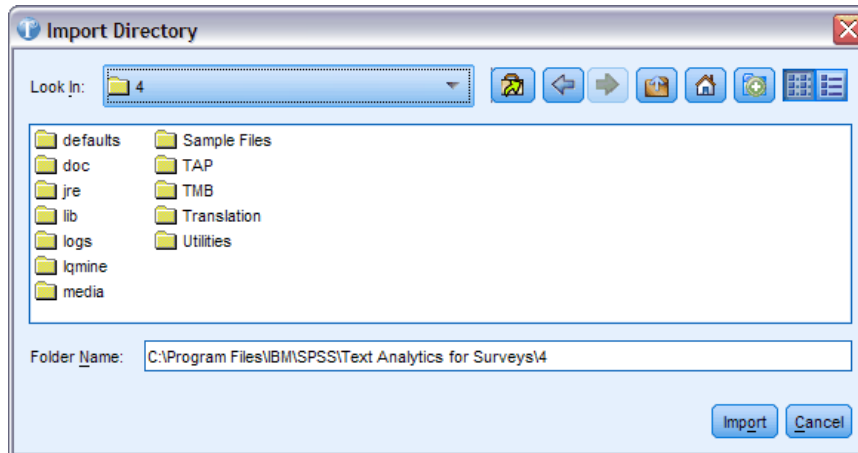


- ▶ Select the file you want to import and click Import. The file contents are transformed into an internal format and added to your library.

To Import All of the Files in a Directory

- ▶ From the menus, choose Resources > Import Files > Import Entire Directory. The Import Directory dialog box opens.

Figure 8-12
Import Directory dialog box



- ▶ Select the library in which you want all of the resource files imported from the Import list. If you select the Default option, a new library will be created using the name of the directory as its name.
- ▶ Select the directory from which to import the files. Subdirectories will not be read.
- ▶ Click Import. The dialog box closes and the content from those imported resource files now appears in the editor in the form of dictionaries and advanced resource files.

Working with Libraries

The resources used by the extraction engine to extract and group terms from your text data always contain one or more libraries. You can see the set of libraries in the library tree located in the upper left part of the Resource Editor. The libraries are composed of three kinds of dictionaries: Type, Substitution, and Exclude. For more information, see the topic “About Library Dictionaries” in Chapter 10 on p. 207.

The resource template or the resources from the TAP you chose includes several libraries to enable you to immediately begin extracting concepts from your text data. However, you can create your own libraries as well and also publish them so you can reuse them. For more information, see the topic “Publishing Libraries” on p. 204.

For example, suppose that you frequently work with text data related to the automotive industry. After analyzing your data, you decide that you would like to create some customized resources to handle industry-specific vocabulary or jargon. Using the Resource Editor, you can create a new template, and in it a library to extract and group automotive terms. Since you will need the information in this library again, you publish your library to a central repository, accessible in the **Manage Libraries** dialog box, so that it can be reused independently in different projects.

Suppose that you are also interested in grouping terms that are specific to different subindustries, such as electronic devices, engines, cooling systems, or even a particular manufacturer or market. You can create a library for each group and then publish the libraries so that they can be used with multiple sets of text data. In this way, you can add the libraries that best correspond to the context of your text data.

Note: Additional resources can be configured and managed in the Advanced Resources tab. Some apply to all of the libraries and manage nonlinguistic entities, fuzzy grouping exceptions, and so on. For more information, see the topic “About Advanced Resources” in Chapter 11 on p. 225.

Shipped Libraries

By default, several libraries are installed with IBM® SPSS® Text Analytics for Surveys. You can use these preformatted libraries to access thousands of predefined terms and synonyms as well as many different types. These shipped libraries are fine-tuned to several different domains and are available in several different languages.

There are a number of libraries but the most commonly used are as follows:

- **Local library.** Used to store user-defined dictionaries. It is an empty library added by default to all resources. It contains an empty type dictionary too. It is most useful when making changes or refinements to the resources directly (such as adding a word to a type) from the text analysis window. In this case, those changes and refinements are automatically stored in the first library listed in the library tree in the Resource Editor; by default, this is the *Local Library*. You cannot publish this library because it is specific to the project data. If you want to publish its contents, you must rename the library first.

- **Core library.** Used in most cases, since it comprises the basic five built-in types representing people, locations, organizations, products, and unknown. While you may see only a few terms listed in one of its type dictionaries, the types represented in the Core library are actually complements to the robust types found in the internal, compiled resources delivered with your text-mining product. These internal, compiled resources contain thousands of terms for each type. For this reason, while you may not see a term in the type dictionary term list, it can still be extracted and typed with a Core type. This explains how names such as *George* can be extracted and typed as <Person> when only *John* appears in the <Person> type dictionary in the Core library. Similarly, if you do not include the Core library, you may still see these types in your extraction results, since the compiled resources containing these types will still be used by the extraction engine.
- **Opinions library.** Used most commonly to extract opinions and sentiments from text data. This library includes thousands of words representing attitudes, qualifiers, and preferences that—when used in conjunction with other terms—indicate an opinion about a subject. This library includes a number of built-in types, synonyms, and excludes. It also includes a large set of pattern rules used for text link analysis.
- **Budget library.** Used to extract terms referring to the cost of something. This library includes many words and phrases that represent adjectives, qualifiers, and judgments regarding the price or quality of something.
- **Variations library.** Used to include cases where certain language variations require synonym definitions to properly group them. This library includes only synonym definitions.

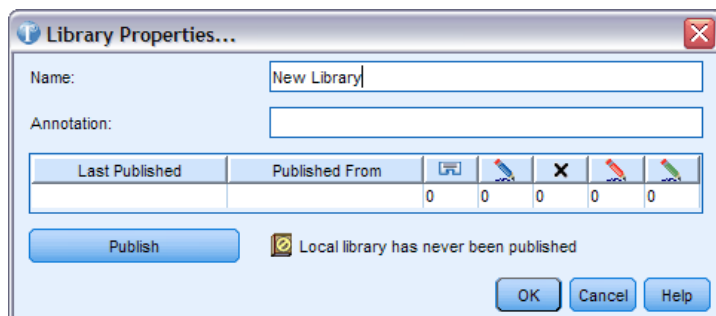
Although some of the libraries shipped outside the templates resemble the contents in some templates, the templates have been specifically tuned to particular applications and contain additional advanced resources. We recommend that you try to use a template that was designed for the kind of text data you are working with and make your changes to those resources rather than just adding individual libraries to a more generic template.

Compiled resources are also delivered with SPSS Text Analytics for Surveys. They are always used during the extraction process and contain a large number of complementary definitions to the built-in type dictionaries in the default libraries. Since these resources are compiled, they cannot be viewed or edited. You can, however, force a term that was typed by these compiled resources into any other dictionary. For more information, see the topic “Forcing Terms” in Chapter 10 on p. 214.

Creating Libraries

You can create any number of libraries. After creating a new library, you can begin to create type dictionaries in this library and enter terms, synonyms, and excludes.

Figure 9-1
Library Properties dialog box



To Create a Library

- ▶ From the menus, choose Resources > New Library. The Add Library to Project dialog opens.
- ▶ Enter a name for the library in the Name text box.
- ▶ If desired, enter a comment in the Annotation text box.
- ▶ Click Publish if you want to publish this library now before entering anything in the library. For more information, see the topic “Sharing Libraries” on p. 202. You can also publish later at any time.
- ▶ Click OK to create the library. The dialog box closes and the library appears in the tree view. If you expand the libraries in the tree, you will see that an empty type dictionary has been automatically included in the library. In it, you can immediately begin adding terms. For more information, see the topic “Adding Terms” in Chapter 10 on p. 210.

Adding Public Libraries

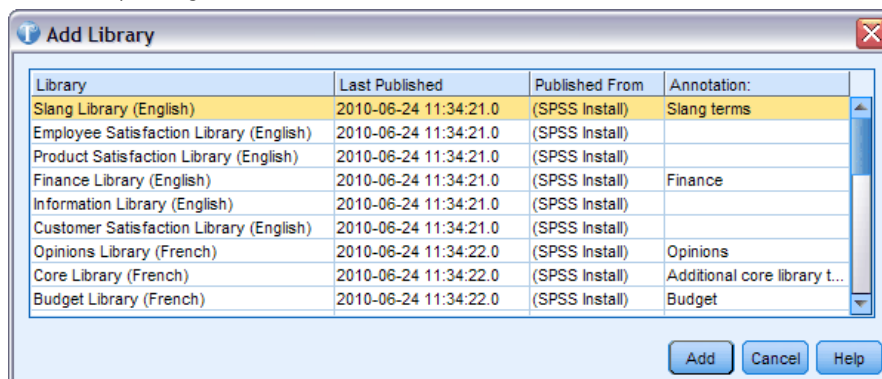
If you want to reuse a library from another project data, you can add it to your current resources as long as it is a public library. A **public library** is a library that has been published. For more information, see the topic “Publishing Libraries” on p. 204.

When you add a public library, a **local** copy is embedded into your project data. You can make changes to this library; however, you must republish the public version of the library if you want to share the changes.

When adding a public library, a Resolve Conflicts dialog box may appear if any conflicts are discovered between the terms and types in one library and the other local libraries. You must resolve these conflicts or accept the proposed resolutions in order to complete this operation. For more information, see the topic “Resolving Conflicts” on p. 205.

Note: If you always update your libraries when you open or publish when you close a project, you are less likely to have libraries that are out of sync. For more information, see the topic “Sharing Libraries” on p. 202.

Figure 9-2
Add Library dialog box



To Add a Library

- ▶ From the menus, choose Resources > Add Library. The Add Library dialog box opens.
- ▶ Select the library or libraries in the list.
- ▶ Click Add. If any conflicts occur between the newly added libraries and any libraries that were already there, you will be asked to verify the conflict resolutions or change them before completing the operation. For more information, see the topic “Resolving Conflicts” on p. 205.

Finding Terms and Types

You can search in the various panes in the editor using the Find feature. In the editor, you can choose Edit > Find from the menus and the Find toolbar appears. You can use this toolbar to find one occurrence at a time. By clicking Find again, you can find subsequent occurrences of your search term.

When searching, the editor searches only the library or libraries listed in the drop-down list on the Find toolbar. If All Libraries is selected, the program will search everything in the editor.

When you start a search, it begins in the area that has the focus. The search continues through each section, looping back around until it returns to the active cell. You can reverse the order of the search using the directional arrows. You can also choose whether or not your search is case sensitive.

To Find Strings in the View

- ▶ From the menus, choose Edit > Find. The Find toolbar appears.
- ▶ Enter the string for which you want to search.
- ▶ Click the Find button to begin the search. The next occurrence of the term or type is then highlighted.
- ▶ Click the button again to move from occurrence to occurrence.

Viewing Libraries

You can display the contents of one particular library or all libraries. This can be helpful when dealing with many libraries or when you want to review the contents of a specific library before publishing it. Changing the view only impacts what you see in this Library Resources tab but does not disable any libraries from being used during extraction. For more information, see the topic “Disabling Local Libraries” on p. 200.

The default view is All Libraries, which shows all libraries in the tree and their contents in other panes. You can change this selection using the drop-down list on the toolbar or through a menu selection (View > Libraries) When a single library is being viewed, all items in other libraries disappear from view but are still read during the extraction.

To Change the Library View

- ▶ From the menus in the Library Resources tab, choose View > Libraries. A menu with all of the local libraries opens.
- ▶ Select the library that you want to see or select the All Libraries option to see the contents of all libraries. The contents of the view are filtered according to your selection.

Managing Local Libraries

Local libraries are the libraries inside your project or inside a template, as opposed to public libraries. For more information, see the topic “Managing Public Libraries” on p. 201. There are also some basic local library management tasks that you might want to perform, including: renaming, disabling, or deleting a local library.

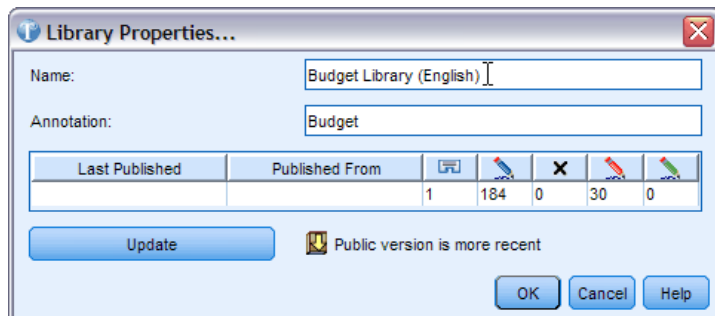
Renaming Local Libraries

You can rename local libraries. If you rename a local library, you will disassociate it from the public version, if a public version exists. This means that subsequent changes can no longer be shared with the public version. You can republish this local library under its new name. This also means that you will not be able to update the original public version with any changes that you make to this local version.

Note: You cannot rename a public library.

- ▶ From the menus, choose Edit > Library Properties. The Library Properties dialog box opens.

Figure 9-3
Library Properties dialog box



To Rename a Local Library

- ▶ In the tree view, select the library that you want to rename.
- ▶ Enter a new name for the library in the Name text box.
- ▶ Click OK to accept the new name for the library. The dialog box closes and the library name is updated in the tree view.

Disabling Local Libraries

If you want to temporarily exclude a library from the extraction process, you can deselect the check box to the left of the library name in the tree view. This signals that you want to keep the library but want the contents ignored when checking for conflicts and during extraction.

To Disable a Library

- ▶ In the library tree pane, select the library you want to disable.
- ▶ Click the spacebar. The check box to the left of the name is cleared.

Deleting Local Libraries

You can remove a library without deleting the public version of the library and vice versa. Deleting a local library will delete the library and all of its content from project only. Deleting a local version of a library does not remove that library from other projects or the public version. For more information, see the topic “Managing Public Libraries” on p. 201.

To Delete a Local Library

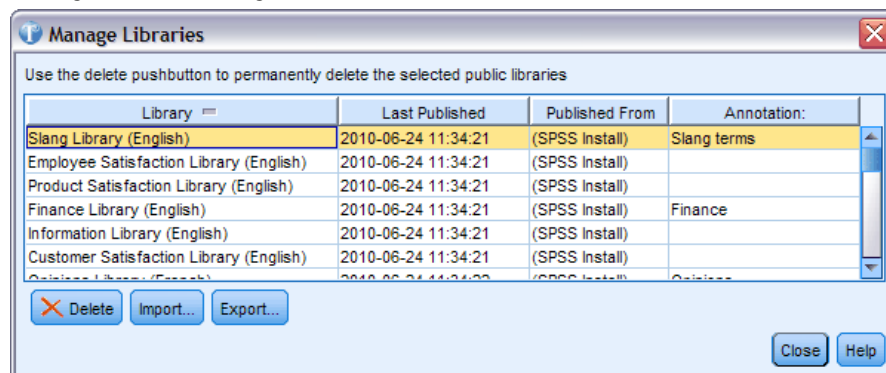
- ▶ In the tree view, select the library you want to delete.
- ▶ From the menus, choose Edit > Delete to delete the library. The library is removed.
- ▶ If you have never published this library before, a message asking whether you would like to delete or keep this library opens. Click Delete to continue or Keep if you would like to keep this library.

Note: One library must always remain.

Managing Public Libraries

In order to reuse local libraries, you can publish them and then work with them and see them through the Manage Libraries dialog box (Resources > Manage Libraries). For more information, see the topic “Sharing Libraries” on p. 202. Some basic public library management tasks that you might want to perform include importing, exporting, or deleting a public library. You cannot rename a public library.

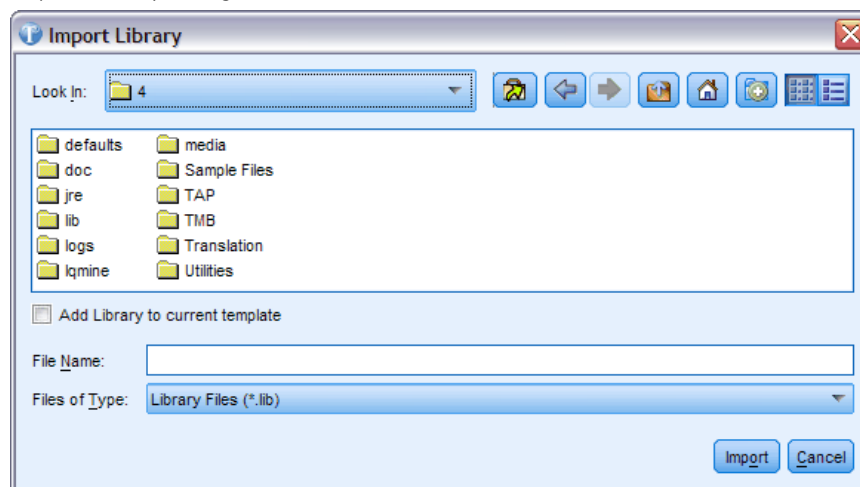
Figure 9-4
Manage Libraries dialog box



Importing Public Libraries

- ▶ In the Manage Libraries dialog box, click Import.... The Import Library dialog box opens.

Figure 9-5
Import Library dialog box



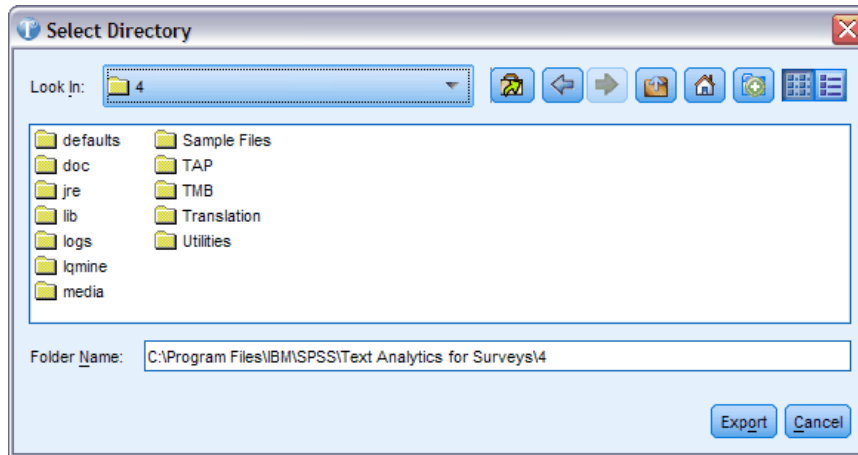
- ▶ Select the library file (*.lib) that you want to import and if you also want to add this library locally, select Add library to current project.
- ▶ Click Import. The dialog box closes. If a public library with the same name already exists, you will be asked to rename the library that you are importing or to overwrite the current public library.

Exporting Public Libraries

You can export public libraries into the *.lib* format so that you can share them.

- ▶ In the Manage Libraries dialog box, select the library that you want to export in the list.
- ▶ Click Export. The Select Directory dialog box opens.

Figure 9-6
Select Directory dialog box



- ▶ Select the directory to which you want to export and click Export. The dialog box closes and the library file (**.lib*) is exported.

Deleting Public Libraries

You can remove a local library without deleting the public version of the library and vice versa. However, if the library is deleted from this dialog box, it can no longer be added to any projects until a local version is published again.

If you delete a library that was installed with the product, the originally installed version is restored.

- ▶ In the Manage Libraries dialog box, select the library that you want to delete. You can sort the list by clicking on the appropriate header.
- ▶ Click Delete to delete the library. IBM® SPSS® Text Analytics for Surveys verifies whether the local version of the library is the same as the public library. If so, the library is removed with no alert. If the library versions differ, an alert opens to ask you whether you want to keep or remove the public version is issued.

Sharing Libraries

Libraries allow you to work with resources in a way that is easy to share among multiple projects. Libraries can exist in two states, or versions. Libraries that are associated with a particular project are called **local libraries**. While working in a project, you may make a lot of changes in the *Vegetables* library, for example. If your changes could be useful with other data, you can make

these resources available by creating a **public library** version of the *Vegetables* library. A public library, as the name implies, is available to any other project.






You can see the public libraries in the Manage Libraries dialog box. Once this public library version exists, you can add it to the resources in other contexts so that these custom linguistic resources can be shared.

The shipped libraries are initially public libraries. It is possible to edit the resources in these libraries and then create a new public version. Those new versions would then be accessible in other new projects.

As you continue to work with your libraries and make changes, your library versions will become desynchronized. In some cases, a local version might be more recent than the public version, and in other cases, the public version might be more recent than the local version. It is also possible for both the public and local versions to contain changes that the other does not if the public version was updated from within another project. If your library versions become desynchronized, you can synchronize them again. Synchronizing library versions consists of republishing and/or updating local libraries.

Whenever you open or close a project, you will be prompted to synchronize any libraries that need updating or republishing. Additionally, you can easily identify the synchronization state of your local library by the icon appearing beside the library name in the tree view or by viewing the Library Properties dialog box. You can also choose to do so at any time through menu selections. The following table describes the five possible states and their associated icons.

Table 9-1
Local library synchronization states

Icon	Local library status description
	Unpublished—The local library has never been published.
	Synchronized—The local and public library versions are identical. This also applies to the <i>Local Library</i> , which cannot be published because it is intended to contain only project-specific resources.
	Out of date—The public library version is more recent than the local version. You can update your local version with the changes.
	Newer—The local library version is more recent than the public version. You can republish your local version to the public version.
	Out of sync—Both the local and public libraries contain changes that the other does not. You must decide whether to update or publish your local library. If you update, you will lose the changes that you made since the last time you updated or published. If you choose to publish, you will overwrite the changes in the public version.

Note: If you always update your libraries when you open or publish when you close a project, you are less likely to have libraries that are out of sync.

You can republish a library any time you think that the changes in the library would benefit other projects that may also contain this library. Then, if your changes would benefit other projects, you can update the local versions in those projects. In this way, you can create projects for each context or domain that applies to your data by creating new libraries and/or adding any number of public libraries to your resources.

If a public version of a library is shared, there is a greater chance that differences between local and public versions will arise. Whenever you open or publish when you close a project, a message appears to enable you to publish and/or update any libraries whose versions are not in sync with those in the Manage Libraries dialog box. If the public library version is more

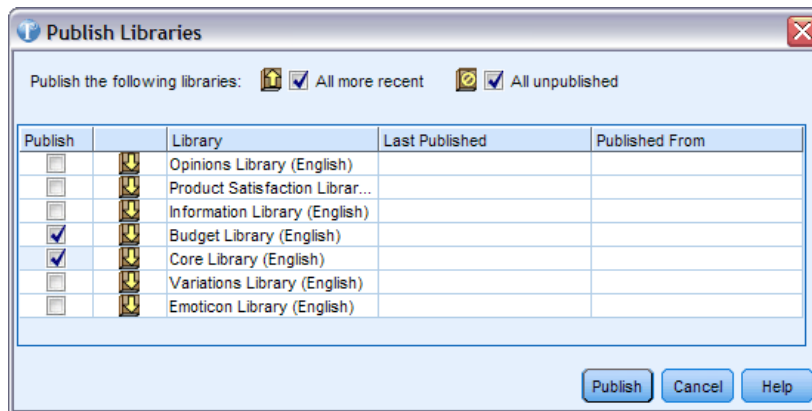
recent than the local version, a dialog box asking whether you would like to update opens. You can choose whether to keep the local version as is instead of updating with the public version or merge the updates into the local library.

Publishing Libraries

If you have never published a particular library, publishing entails creating a public copy of your local library in the database. If you are republishing a library, the contents of the local library will replace the existing public version's contents. After republishing, you can update this library in any other projects so that their local versions are in sync with the public version. Even though you can publish a library, a local version is always stored in the project.

Important! If you make changes to your local library and, in the meantime, the public version of the library was also changed, your library is considered to be out of sync. We recommend that you begin by updating the local version with the public changes, make any changes that you want, and then publish your local version again to make both versions identical. If you make changes and publish first, you will overwrite any changes in the public version.

Figure 9-7
Publish Libraries dialog box



To Publish Local Libraries to the Database

- ▶ From the menus, choose Resources > Publish Libraries. The Publish Libraries dialog box opens, with all libraries in need of publishing selected by default.
- ▶ Select the check box to the left of each library that you want to publish or republish.
- ▶ Click Publish to publish the libraries to the Manage Libraries database.

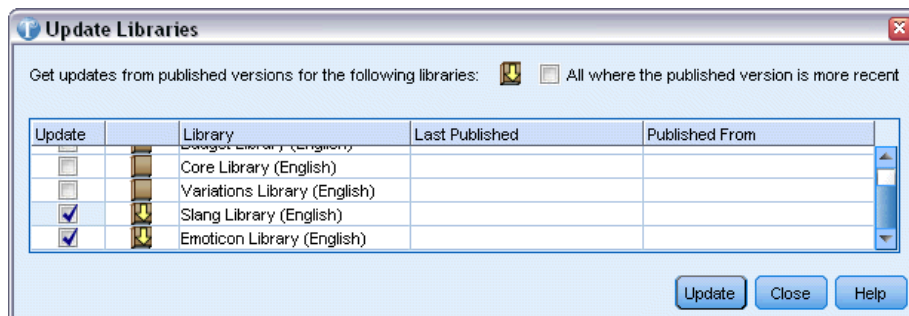
Updating Libraries

Whenever you open or publish when you close a project, you can update or publish any libraries that are no longer in sync with the public versions. If the public library version is more recent than the local version, a dialog box asking whether you would like to update the library opens. You can choose whether to keep the local version instead of updating with the public version or replacing

the local version with the public one. If a public version of a library is more recent than your local version, you can update the local version to synchronize its content with that of the public version. Updating means incorporating the changes found in the public version into your local version.

Note: If you always update your libraries when you open or publish when you close a project, you are less likely to have libraries that are out of sync. For more information, see the topic “Sharing Libraries” on p. 202.

Figure 9-8
Update Libraries dialog box



To Update Local Libraries

- ▶ From the menus, choose Resources > Update Libraries. The Update Libraries dialog box opens, with all libraries in need of updating selected by default.
- ▶ Select the check box to the left of each library that you want to publish or republish.
- ▶ Click Update to update the local libraries.

Resolving Conflicts

Local versus Public Library Conflicts

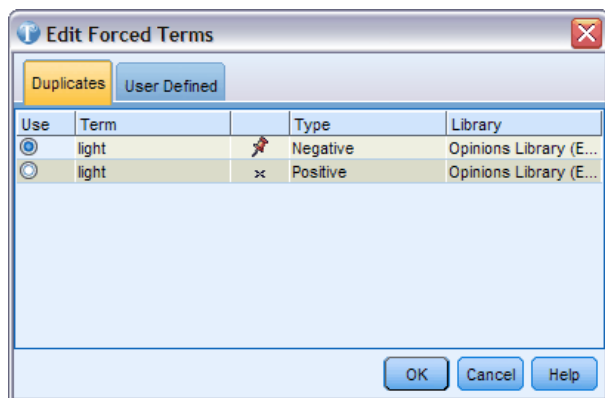
Whenever you open a project, IBM® SPSS® Text Analytics for Surveys performs a comparison of the local libraries and those listed in the Manage Libraries dialog box. If any local libraries in your project are not in sync with the published versions, the Library Synchronization Warning dialog box opens. You can choose from the following options to select the library versions that you want to use here:

- **All libraries local to file.** This option keeps all of your local libraries as they are. You can always republish or update them later.
- **All published libraries on this machine.** This option will replace the shown local libraries with the versions found in the database.
- **All more recent libraries.** This option will replace any older local libraries with the more recent public versions from the database.
- **Other.** This option allows you to manually select the versions that you want by choosing them in the table.

Forced Term Conflicts

Whenever you add a public library or update a local library, conflicts and duplicate entries may be uncovered between the terms and types in this library and the terms and types in the other libraries in your resources. If this occurs, you will be asked to verify the proposed conflict resolutions or change them before completing the operation in the Edit Forced Terms dialog box. For more information, see the topic “Forcing Terms” in Chapter 10 on p. 214.

Figure 9-9
Edit Forced Terms dialog box



The Edit Forced Terms dialog box contains each pair of conflicting terms or types. Alternating background colors are used to visually distinguish each conflict pair. These colors can be changed in the Options dialog box. For more information, see the topic “Options: Display Tab” in Chapter 2 on p. 18. The Edit Forced Terms dialog box contains two tabs:

- **Duplicates.** This tab contains the duplicated terms found in the libraries. If a pushpin icon appears after a term, it means that this occurrence of the term has been forced. If a black X icon appears, it means that this occurrence of the term will be ignored during extraction because it has been forced elsewhere.
- **User Defined.** This tab contains a list of any terms that have been forced manually in the type dictionary term pane and not through conflicts.

Note: The Edit Forced Terms dialog box opens after you add or update a library. If you cancel out of this dialog box, you will not be canceling the update or addition of the library.

To Resolve Conflicts

- ▶ In the Edit Forced Terms dialog box, select the radio button in the Use column for the term that you want to force.
- ▶ When you have finished, click OK to apply the forced terms and close the dialog box. If you click Cancel, you will cancel the changes you made in this dialog box.

About Library Dictionaries

The resources used to extract text data are stored in the form of templates and libraries. A library can be made up of three dictionaries.

- The **type dictionary** contains a collection of terms grouped under one label, or type name. When the extraction engine reads your text data, it compares the words found in the text to the terms defined in your type dictionaries. During extraction, inflected forms of a type's terms and synonyms are grouped under a target term called concept. Extracted concepts are assigned to the type dictionary in which they appear as terms. You can manage your type dictionaries in the upper left and center panes of the editor—the library tree and the term pane. For more information, see the topic “Type Dictionaries” on p. 207.
- The **substitution dictionary** contains a collection of words defined as synonyms or as optional elements used to group similar terms under one target term, called a concept in the final extraction results. You can manage your substitution dictionaries in the lower left pane of the editor using the Synonyms tab and the Optional tab. For more information, see the topic “Substitution/Synonym Dictionaries” on p. 217.
- The **exclude dictionary** contains a collection of terms and types that will be removed from the final extraction results. You can manage your exclude dictionaries in the rightmost pane of the editor. For more information, see the topic “Exclude Dictionaries” on p. 222.

For more information, see the topic “Working with Libraries” in Chapter 9 on p. 195.

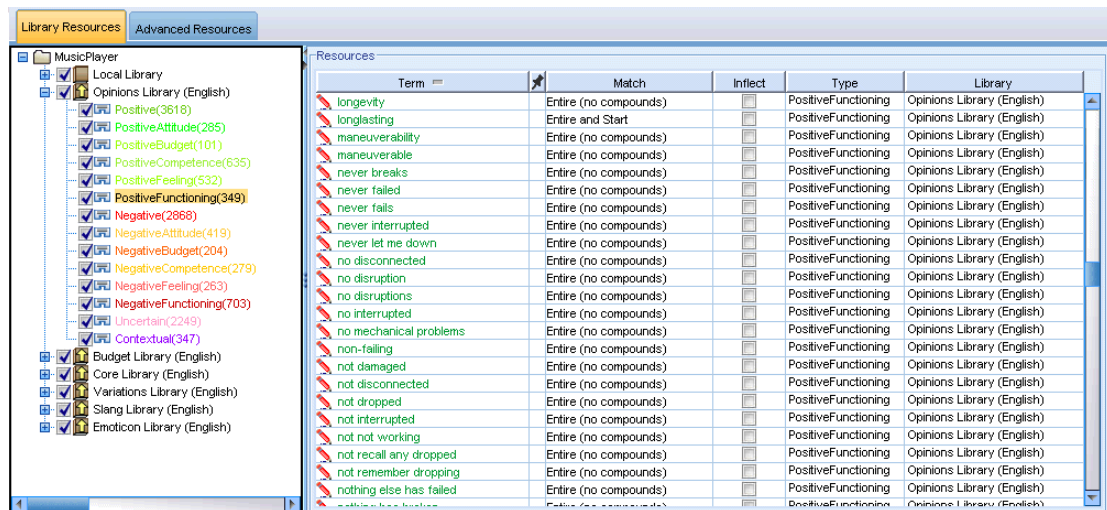
Type Dictionaries

A **type dictionary** is made up of a type name, or label, and a list of terms. Type dictionaries are managed in the upper left and center panes of Library Resources tab in the editor. You can access this view with View > Resource Editor in the menus.

When the extraction engine reads your text data, it compares words found in the text to the terms defined in your type dictionaries. Terms are words or phrases in the type dictionaries in your linguistic resources.

When a word matches a term, it is assigned to the type name for that term. When the resources are read during extraction, the terms that were found in the text then go through several processing steps before they become concepts in the Extraction Results pane. If multiple terms belonging to the same type dictionary are determined to be synonymous by the extraction engine, then they are grouped under the most frequently occurring term and called a **concept** in the Extraction Results pane. For example, if the terms `question` and `query` might appear under the concept name `question` in the end.

Figure 10-1
Library tree and term pane



The list of type dictionaries is shown in the library tree pane on the left. The content of each type dictionary appears in the center pane. Type dictionaries consist of more than just a list of terms. The manner in which words and word phrases in your text data are matched to the terms defined in the type dictionaries is determined by the match option defined. A **match option** specifies how a term is anchored with respect to a candidate word or phrase in the text data. For more information, see the topic “Adding Terms” on p. 210.

Additionally, you can extend the terms in your type dictionary by specifying whether you want to automatically generate and add inflected forms of the terms to the dictionary. By generating the inflected forms, you automatically add plural forms of singular terms, singular forms of plural terms, and adjectives to the type dictionary. For more information, see the topic “Adding Terms” on p. 210.

Note: Concepts that are not found in any type dictionary but are extracted from the text are automatically typed as <Unknown>.

Built-in Types

IBM® SPSS® Text Analytics for Surveys is delivered with a set of linguistic resources in the form of shipped libraries and compiled resources. The shipped libraries contain a set of built-in type dictionaries such as <Location>, <Organization>, <Person>, and <Product>.

These type dictionaries are used by the extraction engine to assign types to the concepts it extracts such as assigned the type <Location> to the concept *paris*. Although a large number of terms have been defined in the built-in type dictionaries, they do not cover every possibility. Therefore, you can add to them or create your own. For a description of the contents of a particular shipped type dictionary, read the annotation in the Type Properties dialog box. Select the type in the tree and choose Edit > Properties from the context menu.

Note: In addition to the shipped libraries, the compiled resources (also used by the extraction engine) contain a large number of definitions complementary to the built-in type dictionaries, but their content is not visible in the product. You can, however, force a term that was typed by the compiled dictionaries into any other dictionary. For more information, see the topic “Forcing Terms” on p. 214.

Creating Types

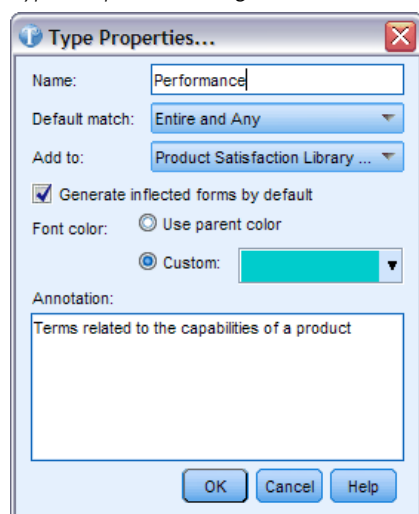
You can create type dictionaries to help group similar terms. When terms appearing in this dictionary are discovered during the extraction process, they will be assigned to this type name and extracted under a concept name. Whenever you create a library, an empty type library is always included so that you can begin entering terms immediately.

If you are analyzing text about food and want to group terms relating to vegetables, you could create your own <Vegetables> type dictionary. You could then add terms such as `carrot`, `broccoli`, and `spinach` if you feel that they are important terms that will appear in the text. Then, during extraction, if any of these terms are found, they are extracted as concepts and assigned to the <Vegetables> type.

You do not have to define every form of a word or expression, because you can choose to generate the inflected forms of terms. By choosing this option, the extraction engine will automatically recognize singular or plural forms of the terms among other forms as belonging to this type. This option is particularly useful when your type contains mostly nouns, since it is unlikely you would want inflected forms of verbs or adjectives.

Important! We strongly recommend that you do not create new types in the Opinions library or else they will not be taken into account during processing. The contents of the Opinions library is handled differently than other libraries since it used to produce patterns. Instead, either work within the types that already exist in that library or add new types to another library in your project.

Figure 10-2
Type Properties dialog box



Name. The name you give to the type dictionary you are creating. We recommend that you do not use spaces in type names, especially if two or more type names start with the same word.

Default match. The default match attribute instructs the extraction engine how to match this term to text data. Whenever you add a term to this type dictionary, this is the match attribute automatically assigned to it. You can always change the match choice manually in the term list. Options include: Entire Term, Start, End, Any, Start or End, Entire and Start, Entire and End, Entire and (Start or End), and Entire (no compounds). For more information, see the topic “Adding Terms” on p. 210.

Add to. This field indicates the library in which you will create your new type dictionary.

Generate inflected forms by default. This option tells the extraction engine to use grammatical morphology to capture and group similar forms of the terms that you add to this dictionary, such as singular or plural forms of the term. This option is particularly useful when your type contains mostly nouns. When you select this option, all new terms added to this type will automatically have this option although you can change it manually in the list.

Font color. This field allows you to distinguish the results from this type from others in the interface. If you select Use parent color, the default type color is used for this type dictionary, as well. This default color is set in the options dialog box. For more information, see the topic “Options: Display Tab” in Chapter 2 on p. 18. If you select Custom, select a color from the drop-down list.

Annotation. This field is optional and can be used for any comments or descriptions.

To Create a Type Dictionary

- ▶ Select the library in which you would like to create a new type dictionary.
- ▶ From the menus, choose Tools > New Type. The Type Properties dialog box opens.
- ▶ Enter the name of your type dictionary in the Name text box and choose the options you want.
- ▶ Click OK to create the type dictionary. The new type is visible in the library tree pane and appears in the center pane. You can begin adding terms immediately. For more information, see “Adding Terms”.

Note: These instructions show you how to make changes within the Resource Editor view. Keep in mind that you can also do this kind of fine-tuning directly from the Extraction Results pane or Data pane. For more information, see the topic “Refining Extraction Results” in Chapter 5 on p. 84.

Adding Terms

The library tree pane displays libraries and can be expanded to show the type dictionaries that they contain. In the center pane, a term list displays the terms in the selected library or type dictionary, depending on the selection in the tree.

Figure 10-3
Term pane

Term	Match	Inflect	Type	Library
non-optimal	Entire Term	<input type="checkbox"/>	NegativeFunctioning	Opinions Library (English)
non-operative	Entire Term	<input type="checkbox"/>	NegativeFunctioning	Opinions Library (English)
non-metallic	Entire Term	<input type="checkbox"/>	Contextual	Opinions Library (English)
non-invasive	Entire Term	<input type="checkbox"/>	PositiveFeeling	Opinions Library (English)
non-intuitive	Entire Term	<input type="checkbox"/>	Negative	Opinions Library (English)
non-intrusive	Entire Term	<input type="checkbox"/>	PositiveFeeling	Opinions Library (English)
non-hostile	Entire Term	<input type="checkbox"/>	PositiveAttitude	Opinions Library (English)
non-functioning	Entire Term	<input type="checkbox"/>	NegativeFunctioning	Opinions Library (English)
non-friendly	Entire Term	<input type="checkbox"/>	NegativeAttitude	Opinions Library (English)
non-fat	Entire Term	<input type="checkbox"/>	Contextual	Opinions Library (English)
non-failing	Entire Term	<input type="checkbox"/>	PositiveFunctioning	Opinions Library (English)
non-existent	Entire Term	<input type="checkbox"/>	Negative	Opinions Library (English)
non-existant	Entire Term	<input type="checkbox"/>	Negative	Opinions Library (English)
non-essential	Entire Term	<input type="checkbox"/>	Negative	Opinions Library (English)
non-equal	Entire Term	<input type="checkbox"/>	Negative	Opinions Library (English)
non-enthusiastic	Entire (no compounds)	<input type="checkbox"/>	NegativeAttitude	Opinions Library (English)
non-enhanced	Entire Term	<input type="checkbox"/>	Negative	Opinions Library (English)
non-enforceable	Entire Term	<input type="checkbox"/>	NegativeFunctioning	Opinions Library (English)
non-effective	Entire Term	<input type="checkbox"/>	Negative	Opinions Library (English)
non-corrected	Entire Term	<input type="checkbox"/>	NegativeCompetence	Opinions Library (English)
non-cooperative	Entire Term	<input type="checkbox"/>	NegativeAttitude	Opinions Library (English)
non-conventional	Entire Term	<input type="checkbox"/>	Positive	Opinions Library (English)
non-constrained	Entire Term	<input type="checkbox"/>	Positive	Opinions Library (English)
non-complexed	Entire Term	<input type="checkbox"/>	Positive	Opinions Library (English)

In the Resource Editor, you can add terms to a type dictionary directly in the term pane or through the Add New Terms dialog box. The terms that you add can be single words or compound words. You will always find a blank row at the top of the list to allow you to add a new term.

Note: These instructions show you how to make changes within the Resource Editor view. Keep in mind that you can also do this kind of fine-tuning directly from the Extraction Results pane or Data pane. For more information, see the topic “Refining Extraction Results” in Chapter 5 on p. 84.

Term Column

In this column, enter single or compound words into the cell. The color in which the term appears depends on the color for the type in which the term is stored or forced. You can change type colors in the Type Properties dialog box. For more information, see the topic “Creating Types” on p. 209.

Force Column

In this column, by putting a pushpin icon into this cell, the extraction engine knows to ignore any other occurrences of this same term in other libraries. For more information, see the topic “Forcing Terms” on p. 214.

Match Column

In this column, select a match option to instruct the extraction engine how to match this term to text data. See the table for examples. You can change the default value by editing the type properties. For more information, see the topic “Creating Types” on p. 209. From the menus,

choose Edit > Change Match. The following are the basic match options since combinations of these are also possible:


- **Start.** If the term in the dictionary matches the beginning of a concept extracted from the text, this type is assigned. For example, if you enter `apple`, `apple tart` will be matched.
- **End.** If the term in the dictionary matches the end of a concept extracted from the text, this type is assigned. For example, if you enter `apple`, `cider apple` will be matched.
- **Any.** If the term in the dictionary matches any part of a concept extracted from the text, this type is assigned. For example, if you enter `apple`, the Any option will type `apple tart`, `cider apple`, and `cider apple tart` the same way.
- **Entire Term.** If the entire concept extracted from the text matches the exact term in the dictionary, this type is assigned. Adding a term as Entire term, Entire and Start, Entire and End, Entire and Any, or Entire (no compounds) will force the extraction of a term.

Furthermore, since the `<Person>` type extracts only two part names, such as *edith piaf* or *mohandas gandhi*, you may want to explicitly add the first names to this type dictionary if you are trying to extract a first name when no last name is mentioned. For example, if you want to catch all instances of *edith* as a name, you should add `edith` to the `<Person>` type using Entire term or Entire and Start.

- **Entire (no compounds).** If the entire concept extracted from the text matches the exact term in the dictionary, this type is assigned and the extraction is stopped to prohibit the extraction from matching the term to a longer compound. For example, if you enter `apple`, the Entire (no compound) option will type `apple` and not extract the compound `apple sauce` unless it is forced in somewhere else.

In the following table, we assume that the term `apple` is in a type dictionary. Depending on the match option, this table shows which concepts would be extracted and typed if they were found in the text.

Table 10-1
Matching Examples

Match options for the term:  <code>apple</code>	Extracted concepts			
	<code>apple</code>	<code>apple tart</code>	<i>ripe apple</i>	<i>homemade apple tart</i>
Entire Term	✓			
Start		✓		
End			✓	
Start or End		✓	✓	
Entire and Start	✓	✓		
Entire and End	✓		✓	
Entire and (Start or End)	✓	✓	✓	
Any		✓	✓	✓
Entire and Any	✓	✓	✓	✓
Entire (no compounds)	✓	<i>never extracted</i>	<i>never extracted</i>	<i>never extracted</i>

Inflect Column

In this column, select whether the extraction engine should generate inflected forms of this term during extraction so that they are all grouped together. The default value for this column is defined in the Type Properties but you can change this option on a case-by-case basis directly in the column. From the menus, choose Edit > Change Inflection.

Type Column

In this column, select a type dictionary from the drop-down list. The list of types is filtered according to your selection in the library tree pane. The first type in the list is always the default type selected in the library tree pane. From the menus, choose Edit > Change Type.

Library Column

In this column, the library in which your term is stored appears. You can drag and drop a term into another type in the library tree pane to change its library.

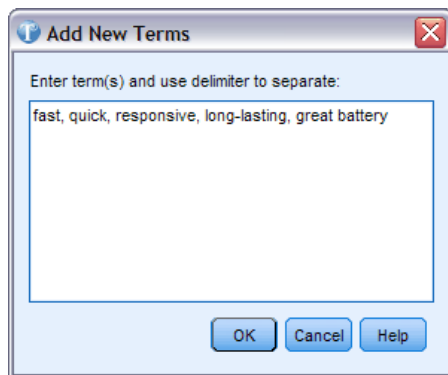
To Add a Single Term to a Type Dictionary

- ▶ In the library tree pane, select the type dictionary to which you want to add the term.
- ▶ In the term list in the center pane, type your term in the first available empty cell and set any options you want for this term.

To Add Multiple Terms to a Type Dictionary

- ▶ In the library tree pane, select the type dictionary to which you want to add terms.
- ▶ From the menus, choose Tools > New Terms. The Add New Terms dialog box opens.

Figure 10-4
Add New Terms dialog box



- ▶ Enter the terms you want to add to the selected type dictionary by typing the terms or copying and pasting a set of terms. If you enter multiple terms, you must separate them using the delimiter that is defined in the Options dialog, or add each term on a new line. For more information, see the topic “Setting Options” in Chapter 2 on p. 16.








- Click OK to add the terms to the dictionary. The match option is automatically set to the default option for this type library. The dialog box closes and the new terms appear in the dictionary.

Forcing Terms

If you want a term to be assigned to a particular type, you can add it to the corresponding type dictionary. However, if there are multiple terms with the same name, the extraction engine must know which type should be used. Therefore, you will be prompted to select which type should be used. This is called **forcing** a term into a type. This option is most useful when overriding the type assignment from a compiled (internal, noneditable) dictionary. In general, we recommend avoiding duplicate terms altogether.

Forcing will *not* remove the other occurrences of this term; rather, they will be ignored by the extraction engine. You can later change which occurrence should be used by forcing or unforcing a term. You may also need to force a term into a type dictionary when you add a public library or update a public library.

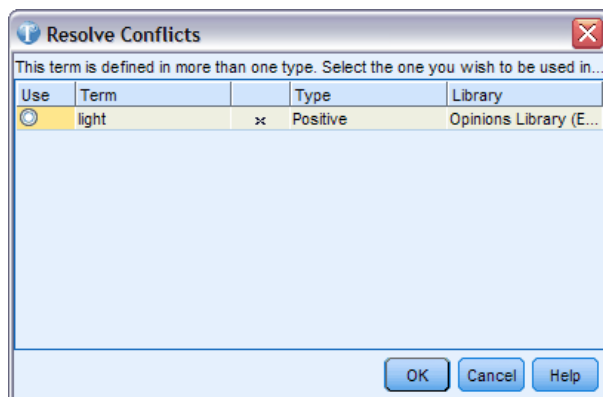
Figure 10-5
Force status icons

Term	Match	Inflect	Type	Library
 bonus	Entire And Any	<input type="checkbox"/>	Budget	Budget Library (English)
 bonuses	Entire And Any	<input type="checkbox"/>	Budget	Budget Library (English)
 bucks	Entire And Any	<input checked="" type="checkbox"/>	Budget	Budget Library (English)
 budget	×× Entire And Any	<input checked="" type="checkbox"/>	Budget	Budget Library (English)
 budget cut backs	Entire Term	<input type="checkbox"/>	Budget	Budget Library (English)
 budget funds	×× Entire Term	<input type="checkbox"/>	Budget	Budget Library (English)
 budget restriction	Entire Term	<input checked="" type="checkbox"/>	Budget	Budget Library (English)

You can see which terms are forced or ignored in the Force column, the second column in the term pane. If a pushpin icon appears, this means that this occurrence of the term has been forced. If a black X icon appears, this means that this occurrence of the term will be ignored during extraction because it has been forced elsewhere. Additionally, when you force a term, it will appear in the color for the type in which it was forced. This means that if you forced a term that is in both `Type 1` and `Type 2` into `Type 1`, any time you see this term in the window, it will appear in the font color defined for `Type 1`.

You can double-click the icon in order to change the status. If the term appears elsewhere, a Resolve Conflicts dialog box opens to allow you to select which occurrence should be used.

Figure 10-6
Resolve Conflicts dialog box



Renaming Types

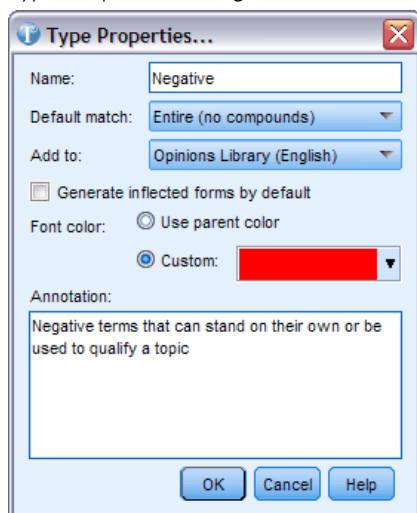
You can rename a type dictionary or change other dictionary settings by editing the type properties.

Important! We recommend that you do not use spaces in type names, especially if two or more type names start with the same word. We also recommend that you do not rename the types in the Core or Opinions libraries or change their default match attributes.

To Rename a Type

- ▶ In the library tree pane, select the type dictionary you want to rename.
- ▶ Right-click your mouse and choose Type Properties from the context menu. The Type Properties dialog box opens.

Figure 10-7
Type Properties dialog box



- ▶ Enter the new name for your type dictionary in the Name text box.

- ▶ Click OK to accept the new name. The new type name is visible in the library tree pane.

Moving Types

You can drag a type dictionary to another location within a library or to another library in the tree.

Note: We recommend that you do not move the built-in types.

To Reorder a Type within a Library

- ▶ In the library tree pane, select the type dictionary you want to move.
- ▶ From the menus, choose Edit > Move Up to move the type dictionary up one position in the library tree pane or Edit > Move Down to move it down one position.

To Move a Type to Another Library

- ▶ In the library tree pane, select the type dictionary you want to move.
- ▶ Right-click your mouse and choose Type Properties from the context menu. The Type Properties dialog box opens. (You can also drag and drop the type into another library).
- ▶ In the Add To list box, select the library to which you want to move the type dictionary.
- ▶ Click OK. The dialog box closes, and the type is now in the library you selected.

Disabling and Deleting Types

If you want to temporarily remove a type dictionary, you can disable it by deselecting the check box to the left of the dictionary name in the library tree pane. This signals that you want to keep the dictionary in your library but want the contents ignored during conflict checking and during the extraction process.

You can also permanently delete type dictionaries from a library.

Note: We recommend that you do not delete the built-in types in the Core or Opinions libraries. We recommend disabling them instead.

To Disable a Type Dictionary

- ▶ In the library tree pane, select the type dictionary you want to disable.
- ▶ Click the spacebar. The check box to the left of the type name is cleared.

To Delete a Type Dictionary

- ▶ In the library tree pane, select the type dictionary you want to delete.
- ▶ From the menus, choose Edit > Delete to delete the type dictionary.

Substitution/Synonym Dictionaries

A **substitution dictionary** is a collection of terms that help to group similar terms under one target term. Substitution dictionaries are managed in the bottom pane of the Library Resources tab. You can access this view with View > Resource Editor in the menus.

You can define two forms of substitutions in this dictionary: **synonyms** and **optional elements**. You can click the tabs in this pane to switch between them.

After you run an extraction on your text data, you may find several concepts that are synonyms or inflected forms of other concepts. By identifying optional elements and synonyms, you can force the extraction engine to map these to one single target term.

Substituting using synonyms and optional elements reduces the number of concepts in the Extraction Results pane by combining them together into more significant, representative concepts with higher frequency counts.

Figure 10-8
Substitution dictionary pane

	Target	Synonyms	Library
0			
1	look	look, lookin, the way it looks	Product Satisfaction Library (English)
2	advertisement	ad, advert, advertasing, advertise, advertising, advertisement	Product Satisfaction Library (English)
3	aftertaste	after taste, after-taste	Product Satisfaction Library (English)
4	anti-spam	anti spam, antispam	Product Satisfaction Library (English)
5	appearance	appearance	Product Satisfaction Library (English)
6	authorisation	authorise, authorising, authorization, authorize, authorizing	Product Satisfaction Library (English)
7	battery	abibtery, batery, batt, battery life, batttery	Product Satisfaction Library (English)
8	call-waiting	call waiting	Product Satisfaction Library (English)
9	characteristic	attribute, charatceristic, properties	Product Satisfaction Library (English)
10	comfort	confort	Product Satisfaction Library (English)
	communication	^ communicate, ^ communicate, amount of mail, commuication, communciation,	Product Satisfaction Library (English)

Synonyms Optional Elements

Synonyms

Synonyms associate two or more words that have the same meaning. You can also use synonyms to group terms with their abbreviations or to group commonly misspelled words with the correct spelling. You can define these synonyms on the Synonyms tab.

A synonym definition is made up of two parts. The first is a Target term, which is the term under which you want the extraction engine to group all synonym terms. Unless this target term is used as a synonym of another target term or unless it is excluded, it is likely to become the concept that appears in the Extraction Results pane. The second is the list of synonyms that will be grouped under the target term.

For example, if you want `automobile` to be replaced by `vehicle`, then `automobile` is the synonym and `vehicle` is the target term.

You can enter any words into the Synonym column, but if the word is not found during extraction and the term had a match option with `Entire`, then no substitution can take place. However, the target term does not need to be extracted for the synonyms to be grouped under this term.

Figure 10-9
Substitution dictionary, Synonyms tab

	Target	Synonyms	Library
1	<input checked="" type="checkbox"/> vehicle	automobile	Local Library
2	<input checked="" type="checkbox"/> look	look, lookin, the way it looks	Product Satisfaction Library
3	<input checked="" type="checkbox"/> advertisement	ad, advert, advertasing, advertise, advertising, advertisement	Product Satisfaction Library
4	<input checked="" type="checkbox"/> aftertaste	after taste, after-taste	Product Satisfaction Library
5	<input checked="" type="checkbox"/> anti-spam	anti spam, antispam	Product Satisfaction Library
6	<input checked="" type="checkbox"/> appearance	appearence	Product Satisfaction Library
7	<input checked="" type="checkbox"/> authorisation	authorise, authorising, authorization, authorize, authorizing	Product Satisfaction Library

Synonyms Optional Elements

Optional Elements

Optional elements identify optional words in a compound term that can be ignored during extraction in order to keep similar terms together even if they appear slightly different in the text. Optional elements are single words that, if removed from a compound, could create a match with another term. These single words can appear anywhere within the compound—at the beginning, middle, or end. You can define optional elements on the Optional tab.

For example, to group the terms `ibm` and `ibm corp` together, you should declare `corp` to be treated as an optional element in this case. In another example, if you designate the term `access` to be an optional element and during extraction both `internet access speed` and `internet speed` are found, they will be grouped together under the term that occurs most frequently.

Figure 10-10
Substitution dictionary, Optional tab

	Optional Elements	Library
<input checked="" type="checkbox"/>		Local Library
<input checked="" type="checkbox"/>		Product Satisfaction Library (E
<input checked="" type="checkbox"/>		Opinions Library (English)
<input checked="" type="checkbox"/>		Budget Library (English)
<input checked="" type="checkbox"/>	a.g., a.g., ag, co., co., corp, corp., corporation, gbh, gmbh, inc, inc., incorporated, kga, l.l.c., l.l.c., llc, ltd, ltd., org, plc, s.a., s.a., s.c.a., s.c.a., sa, sca	Core Library (English)
<input checked="" type="checkbox"/>		Variations Library (English)

Synonyms Optional Elements

Defining Synonyms

On the Synonyms tab, you can enter a synonym definition in the empty line at the top of the table. Begin by defining the target term and its synonyms. You can also select the library in which you would like to store this definition. During extraction, all occurrences of the synonyms will be grouped under the target term in the final extraction. For more information, see the topic “Adding Terms” on p. 210.

For example, if your text data includes a lot of telecommunications information, you may have these terms: `cellular phone`, `wireless phone`, and `mobile phone`. In this example, you may want to define `cellular` and `mobile` as synonyms of `wireless`. If you define these synonyms, then every extracted occurrence of `cellular phone` and `mobile phone` will be treated as the same term as `wireless phone` and will appear together in the term list.

When you are building your type dictionaries, you may enter a term and then think of three or four synonyms for that term. In that case, you could enter all of the terms and then your target term into the substitution dictionary and then drag the synonyms.

Synonym substitution is also applied to the inflected forms (such as the plural form) of the synonym. Depending on the context, you may want to impose constraints on how terms are substituted. Certain characters can be used to place limits on how far the synonym processing should go:

- **Exclamation mark (!).** When the exclamation mark directly precedes the synonym `!synonym`, this indicates that no inflected forms of the synonym will be substituted by the target term. However, an exclamation mark directly preceding the target term `!target-term` means that you do not want any part of the compound target term or variants to receive any further substitutions.
- **Asterisk (*).** An asterisk placed directly after a synonym, such as `synonym*`, means that you want this word to be replaced by the target term. For example, if you defined `manage*` as the synonym and `management` as the target, then `associate managers` will be replaced by the target term `associate management`. You can also add a space and an asterisk after the word (`synonym *`) such as `internet *`. If you defined the target as `internet` and the synonyms as `internet * *` and `web *`, then `internet access card` and `web portal` would be replaced with `internet`. You cannot begin a word or string with the asterisk wildcard in this dictionary.
- **Caret (^).** A caret and a space preceding the synonym, such as `^ synonym`, means that the synonym grouping applies only when the term begins with the synonym. For example, if you define `^ wage` as the synonym and `income` as the target and both terms are extracted, then they will be grouped together under the term `income`. However, if `minimum wage` and `income` are extracted, they will not be grouped together, since `minimum wage` does not begin with `wage`. A space must be placed between this symbol and the synonym.
- **Dollar sign (\$).** A space and a dollar sign following the synonym, such as `synonym $`, means that the synonym grouping applies only when the term ends with the synonym. For example, if you define `cash $` as the synonym and `money` as the target and both terms are extracted, then they will be grouped together under the term `money`. However, if `cash cow` and `money` are extracted, they will not be grouped together, since `cash cow` does not end with `cash`. A space must be placed between this symbol and the synonym.
- **Caret (^) and dollar sign (\$).** If the caret and dollar sign are used together, such as `^ synonym $`, a term matches the synonym only if it is an exact match. This means that no words can appear before or after the synonym in the extracted term in order for the synonym grouping to take place. For example, you may want to define `^ van $` as the synonym and `truck` as the target so that only `van` is grouped with `truck`, while `marie van guerin` will be left unchanged. Additionally, whenever you define a synonym using the caret and dollar signs and this word appears anywhere in the source text, the synonym is automatically extracted.

Figure 10-11
Substitution dictionary, Synonyms tab with example

	Target	Synonyms	Library
1	vehicle	automobile	Local Library
2	look	look, lookin, the way it looks	Product Satisfaction Library
3	advertisement	ad, advert, advertasing, advertise, advertising, advertisement	Product Satisfaction Library
4	aftertaste	after taste, after-taste	Product Satisfaction Library
5	anti-spam	anti spam, antispan	Product Satisfaction Library
6	appearance	appearence	Product Satisfaction Library
7	authorisation	authorise, authorising, authorization, authorize, authorizing	Product Satisfaction Library

Synonyms Optional Elements

To Add a Synonym Entry

- ▶ With the substitution pane displayed, click the Synonyms tab in the lower left corner.
- ▶ In the empty line at the top of the table, enter your target term in the Target column. The target term you entered appears in color. This color represents the type in which the term appears or is forced, if that is the case. If the term appears in black, this means that it does not appear in any type dictionaries.
- ▶ Click in the second cell to the right of the target and enter the set of synonyms. Separate each entry using the global delimiter as defined in the Options dialog box. For more information, see the topic “Setting Options” in Chapter 2 on p. 16. The terms that you enter appear in color. This color represents the type in which the term appears. If the term appears in black, this means that it does not appear in any type dictionaries.
- ▶ Click in the last cell to select the library in which you want to store this synonym definition.

Note: These instructions show you how to make changes within the Resource Editor view. Keep in mind that you can also do this kind of fine-tuning directly from the Extraction Results pane or Data pane. For more information, see the topic “Refining Extraction Results” in Chapter 5 on p. 84.

Defining Optional Elements

On the Optional tab, you can define optional elements for any library you want. These entries are grouped together for each library. As soon as a library is added to the library tree pane, an empty optional element line is added to the Optional tab.

All entries are transformed into lowercase words automatically. The extraction engine will match entries to both lowercase and uppercase words in the text.

Figure 10-12
Substitution dictionary, Optional tab

Optional Elements	Library
<input checked="" type="checkbox"/>	Local Library
<input checked="" type="checkbox"/>	Product Satisfaction Library (E
<input checked="" type="checkbox"/>	Opinions Library (English)
<input checked="" type="checkbox"/>	Budget Library (English)
<input checked="" type="checkbox"/> a.g., a.g., ag, co, co., corp, corp., corporation, gbh, gmbh, inc, inc., incorporated, kga, l.l.c., l.l.c., llc, ltd, ltd., org, plc, s.a, s.a., s.c.a, s.c.a., sa, sca	Core Library (English)
<input checked="" type="checkbox"/>	Variations Library (English)

Synonyms Optional Elements

Note: Terms are delimited using the delimiter defined in the Options dialog. For more information, see the topic “Setting Options” in Chapter 2 on p. 16. If the optional element that you are entering includes the same delimiter as part of the term, a backslash must precede it.

To Add an Entry

- ▶ With the substitution pane displayed, click the Optional tab in the lower left corner of the editor.
- ▶ Click in the cell in the Optional Elements column for the library to which you want to add this entry.
- ▶ Enter the optional element. Separate each entry using the global delimiter as defined in the Options dialog box. For more information, see the topic “Setting Options” in Chapter 2 on p. 16.

Disabling and Deleting Substitutions

You can remove an entry in a temporary manner by disabling it in your dictionary. By disabling an entry, the entry will be ignored during extraction.

You can also delete any obsolete entries in your substitution dictionary.

To Disable an Entry

- ▶ In your dictionary, select the entry you want to disable.
- ▶ Click the spacebar. The check box to the left of the entry is cleared.

Note: You can also deselect the check box to the left of the entry to disable it.

To Delete a Synonym Entry

- ▶ In your dictionary, select the entry you want to delete.
- ▶ From the menus, choose Edit > Delete or press the Delete key on your keyboard. The entry is no longer in the dictionary.

To Delete an Optional Element Entry

- ▶ In your dictionary, double-click the entry you want to delete.

- ▶ Manually delete the term.
- ▶ Press Enter to apply the change.

Exclude Dictionaries

An **exclude dictionary** is a list of words, phrases, or partial strings. Any terms matching or containing an entry in the exclude dictionary will be ignored or excluded from extraction. Exclude dictionaries are managed in the right pane of the editor. Typically, the terms that you add to this list are fill-in words or phrases that are used in the text for continuity but that do not really add anything important to the text and may clutter the extraction results. By adding these terms to the exclude dictionary, you can make sure that they are never extracted.

Exclude dictionaries are managed in the upper right pane of Library Resources tab in the editor. You can access this view with View > Resource Editor in the menus.

Figure 10-13
Exclude dictionary pane

	Exclude List	Library
0	<input type="checkbox"/>	
1	<input checked="" type="checkbox"/> any kind of problem	Opinions Library (English)
2	<input checked="" type="checkbox"/> any problems i have	Opinions Library (English)
3	<input checked="" type="checkbox"/> anykind of problem	Opinions Library (English)
4	<input checked="" type="checkbox"/> as usual	Opinions Library (English)
5	<input checked="" type="checkbox"/> can't wait	Opinions Library (English)
6	<input checked="" type="checkbox"/> i was out of	Opinions Library (English)
7	<input checked="" type="checkbox"/> if i ever have a problem	Opinions Library (English)
8	<input checked="" type="checkbox"/> if i ever have problems	Opinions Library (English)
9	<input checked="" type="checkbox"/> if i have a problem	Opinions Library (English)
10	<input checked="" type="checkbox"/> if i have questions	Opinions Library (English)
11	<input checked="" type="checkbox"/> if there are problems	Opinions Library (English)
12	<input checked="" type="checkbox"/> if there is a problem	Opinions Library (English)

In the exclude dictionary, you can enter a word, phrase, or partial string in the empty line at the top of the table. You can add character strings to your exclude dictionary as one or more words or even partial words using the asterisk as a wildcard. The entries declared in the exclude dictionary will be used to bar concepts from extraction. If an entry is also declared somewhere else in the interface, such as in a type dictionary, it is shown with a strike-through in the other dictionaries, indicating that it is currently excluded. This string does not have to appear in the text data or be declared as part of any type dictionary to be applied.

Note: If you add a concept to the exclude dictionary that also acts as the target in a synonym entry, then the target and all of its synonyms will also be excluded. For more information, see the topic “Defining Synonyms” on p. 218.

Using Wildcards (*)

You can use the asterisk wildcard to denote that you want the exclude entry to be treated as a partial string. Any terms found by the extraction engine that contain a word that begins or ends with a string entered in the exclude dictionary will be excluded from the final extraction. However, there are two cases where the wildcard usage is not permitted:

- Dash character (-) preceded by an asterisk wildcard, such as *-
- Apostrophe (') preceded by an asterisk wildcard, such as *'

Table 10-2
Examples of exclude entries

Entry	Example	Results
word	<i>next</i>	No concepts (or its terms) will be extracted if they contain the word <i>next</i> .
phrase	<i>for example</i>	No concepts (or its terms) will be extracted if they contain the phrase <i>for example</i> .
partial	<i>copyright*</i>	Will exclude any concepts (or its terms) matching or containing the variations of the word <i>copyright</i> , such as <i>copyrighted</i> , <i>copyrighting</i> , <i>copyrights</i> , or <i>copyright 2010</i> .
partial	<i>*ware</i>	Will exclude any concepts (or its terms) matching or containing the variations of the word <i>ware</i> , such as <i>freeware</i> , <i>shareware</i> , <i>software</i> , <i>hardware</i> , <i>beware</i> , or <i>silverware</i> .

To Add Entries

- ▶ In the empty line at the top of the table, enter a term. The term that you enter appears in color. This color represents the type in which the term appears. If the term appears in black, this means that it does not appear in any type dictionaries.

To Disable Entries

You can temporarily remove an entry by disabling it in your exclude dictionary. By disabling an entry, the entry will be ignored during extraction.

- ▶ In your exclude dictionary, select the entry that you want to disable.
- ▶ Click the spacebar. The check box to the left of the entry is cleared.

Note: You can also deselect the check box to the left of the entry to disable it.

To Delete Entries

You can delete any unneeded entries in your exclude dictionary.

- ▶ In your exclude dictionary, select the entry that you want to delete.
- ▶ From the menus, choose Edit > Delete. The entry is no longer in the dictionary.

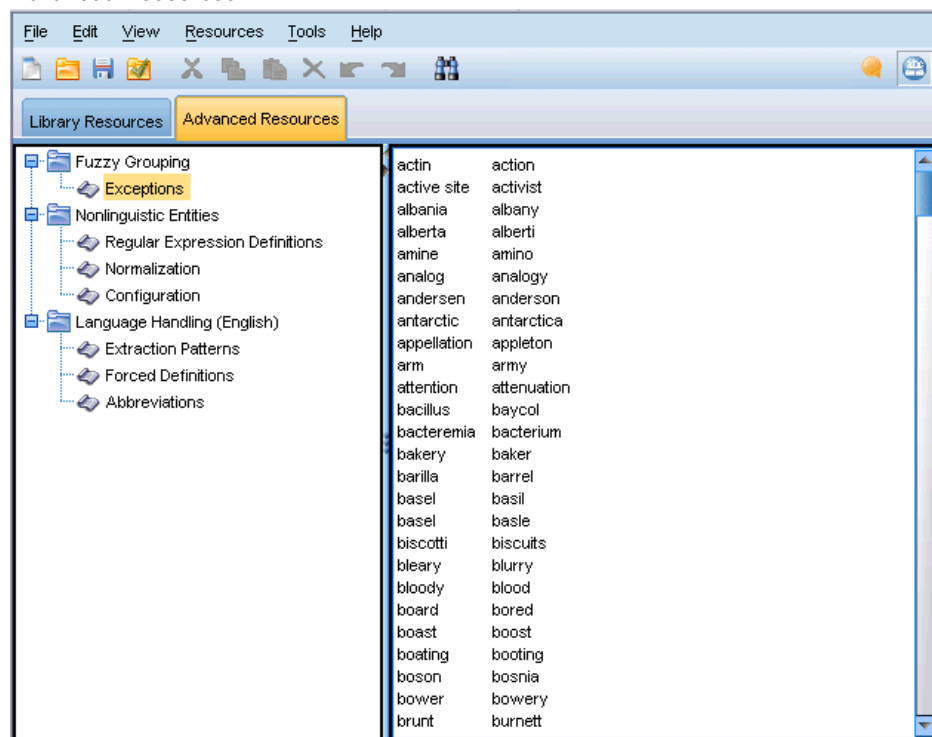
About Advanced Resources

In addition to type, exclude and substitution dictionaries, you can also work with a variety of advanced resource settings such as Fuzzy Grouping settings or nonlinguistic type definitions. You can work with these resources in the Advanced Resources tab in the Resource Editor view. You can also save your changes as the default for all projects, or you can revert back to the original content.

When you go to the Advanced Resources tab, you can edit the following information:

- **Fuzzy Grouping (Exceptions).** Used to exclude word pairs from the fuzzy grouping (spelling error correction) algorithm. For more information, see the topic “Fuzzy Grouping” on p. 227.
- **Nonlinguistic Entities.** Used to enable and disable which nonlinguistic entities can be extracted, as well as the regular expressions and the normalization rules that are applied during their extraction. For more information, see the topic “Nonlinguistic Entities” on p. 228.
- **Language Handling.** Used to declare the special ways of structuring sentences (extraction patterns and forced definitions) and using abbreviations for the selected language. For more information, see the topic “Language Handling” on p. 233.

Figure 11-1
Advanced Resources



Note: You can use the Find/Replace toolbar to find information quickly or to make uniform changes to a section. For more information, see the topic “Replacing” on p. 226.

To Edit Advanced Resources

- ▶ Locate and select the resource section that you want to edit. The contents appear in the right pane.
- ▶ Use the menu or the toolbar buttons to cut, copy, or paste content, if necessary.
- ▶ Edit the file(s) that you want to change using the formatting rules in this section. Your changes are saved as soon as you make them. Use the undo or redo arrows on the toolbar to revert to the previous changes.

Finding

In some cases, you may need to locate information quickly in a particular section. Using the Find feature, you can find a specific rule quickly. To search for information in a section, you can use the Find toolbar.

Figure 11-2
Find toolbar

**To Use the Find Feature**

- ▶ Locate and select the resource section that you want to search. The contents appear in the right pane of the editor.
- ▶ From the menus, choose Edit > Find. The Find toolbar appears at the upper right of the Edit Advanced Resources dialog box.
- ▶ Enter the word string that you want to search for in the text box. You can use the toolbar buttons to control the case sensitivity, partial matching, and direction of the search.
- ▶ Click Find to start the search. If a match is found, the text is highlighted in the window.
- ▶ Click Find again to look for the next match.

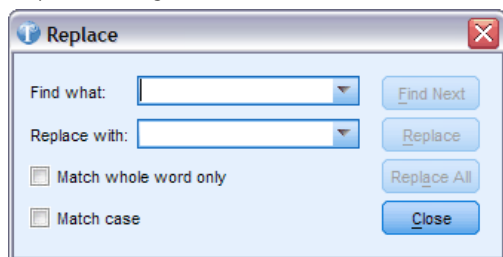
Replacing

In some cases, you may need to make broader updates to your advanced resources. The Replace feature can help you to make uniform updates to your content.

To Use the Replace Feature

- ▶ Locate and select the resource section in which you want to search and replace. The contents appear in the right pane of the editor.
- ▶ From the menus, choose Edit > Replace. The Replace dialog box opens.

Figure 11-3
Replace dialog box



- ▶ In the Find what text box, enter the word string that you want to search for.
- ▶ In the Replace with text box, enter the string that you want to use in place of the text that was found.
- ▶ Select Match whole word only if you want to find or replace only complete words.
- ▶ Select Match case if you want to find or replace only words that match the case exactly.
- ▶ Click Find Next to find a match. If a match is found, the text is highlighted in the window. If you do not want to replace this match, click Find Next again until you find a match that you want to replace.
- ▶ Click Replace to replace the selected match.
- ▶ Click Replace to replace all matches in the section. A message opens with the number of replacements made.
- ▶ When you are finished making your replacements, click Close. The dialog box closes.

Note: If you made a replacement error, you can undo the replacement by closing the dialog box and choosing Edit > Undo from the menus. You must perform this once for every change that you want to undo.

Fuzzy Grouping

In the Extraction Settings dialog, if you select Accommodate spelling for a minimum root character limit of:, you have enabled the fuzzy grouping algorithm.

Fuzzy grouping helps to group commonly misspelled words or closely spelled words by temporarily stripping all vowels (except for the first vowel) and double or triple consonants from extracted words and then comparing them to see if they are the same. During the extraction process, the fuzzy grouping feature is applied to the extracted terms and the results are compared to determine whether any matches are found. If so, the original terms are grouped together in the final extraction list. They are grouped under the term that occurs most frequently in the data.

Note: If the two terms being compared are assigned to different types, excluding the <Unknown> type, then the fuzzy grouping technique is not be applied to this pair. In other words, the terms must belong to the same type or the <Unknown> type in order for the technique to be applied.

If you enabled this feature and found that two words with similar spelling were incorrectly grouped together, you may want to exclude them from fuzzy grouping. You can do this by entering the incorrectly matched pairs into the Exceptions section in the Advanced Resources tab. For more information, see the topic “About Advanced Resources” on p. 225.

The following example demonstrates how fuzzy grouping is performed. If fuzzy grouping is enabled, these words appear to be the same and are matched in the following manner:

```

color -> colr           mountain -> montn
colour -> colr         montana -> montn

modeling -> modlng     furniture -> furntr
modelling -> modlng    furnature -> furntr

```

In the preceding example, you would most likely want to exclude `mountain` and `montana` from being grouped together. Therefore, you could enter them in the Exceptions section in the following manner:

```

mountain      montana

```

Important! In some cases, fuzzy grouping exceptions do not stop 2 words from being paired because certain synonym rules are being applied. In that case, you may want to try entering synonyms using the exclamation mark wildcard (!) to prohibit the words from becoming synonymous in the output. For more information, see the topic “Defining Synonyms” in Chapter 10 on p. 218.

Formatting Rules for Fuzzy Grouping Exceptions

- Define only one exception pair per line.
- Use simple or compound words.
- Use only lowercase characters for the words. Uppercase words will be ignored.
- Use a TAB character to separate each word in a pair.

Nonlinguistic Entities

When working with certain kinds of data, you might be very interested in extracting dates, social security numbers, percentages, or other nonlinguistic entities. These entities are explicitly declared in the configuration file, in which you can enable or disable the entities. For more information, see the topic “Configuration” on p. 232. In order to optimize the output from the extraction engine, the input from nonlinguistic processing is normalized to group like entities according to predefined formats. For more information, see the topic “Normalization” on p. 231.

Note: You can turn on and off nonlinguistic entity extraction in the extraction settings.

Available Nonlinguistic Entities

The nonlinguistic entities in the following table can be extracted. The type name is in parentheses.

Addresses (<Address>)	Organizations (<Organization>)
Amino acids (<Aminoacid>)	Percentages (<Percent>)
Currencies (<Currency>)	Products (<Product>)
Dates (<Date>)	Proteins (<Gene>)
Delay (<Delay>)	Phone numbers (<PhoneNumber>)
Digits (<Digit>)	Times (<Time>)

E-mail addresses (<email>)	U.S. social security (<SocialSecurityNumber>)
HTTP/URL addresses (<url>)	Weights and measures (<Weights-Measures>)
IP address (<IP>)	

Cleaning Text for Processing

Before nonlinguistic entities extraction occurs, the input text is cleaned. During this step, the following temporary changes are made so that nonlinguistic entities can be identified and extracted as such:

- Any sequence of two or more spaces is replaced by a single space.
- Tabulations are replaced by space.
- Single end-of-line characters or sequence characters are replaced by a space, while multiple end-of-line sequences are marked as end of a paragraph. End of line can be denoted by carriage returns (CR) and line feed (LF) or even both together.
- HTML and XML tags are temporarily stripped and ignored.

Regular Expression Definitions

When extracting nonlinguistic entities, you may want to edit or add to the regular expression definitions that are used to identify regular expressions. This is done in the Regular Expression Definitions section in the Advanced Resources tab. For more information, see the topic “About Advanced Resources” on p. 225.

The file is broken up into distinct sections. The first section is called [macros]. In addition to that section, an additional section can exist for each nonlinguistic entity. You can add sections to this file. Within each section, rules are numbered (*regex1*, *regex2*, and so on). These rules must be numbered sequentially from 1–*n*. Any break in numbering will cause the processing of this file to be suspended altogether.

In certain cases, an entity is language dependent. An entity is considered to be language dependent if it takes a value other than 0 for the language parameter in the configuration file. For more information, see the topic “Configuration” on p. 232. When an entity is language dependent, the language must be used to prefix the section name, such as [english/PhoneNumber]. That section would contain rules that apply only to English phone numbers when the PhoneNumber entity is given a value of 2 for the language.

Important! If you make changes to this file or any other in the editor and the extraction engine no longer works as desired, use the Reset to Original option on the toolbar to reset the file to the original shipped content. This file requires a certain level of familiarity with regular expressions. If you require additional assistance in this area, please contact IBM Corp. for help.

Special Characters . [] {} () \ * + ? | ^ \$

All characters match themselves except for the following special characters, which are used for a specific purpose in expressions: . [{ () \ * + ? | ^ \$ To use these characters as such, they must be preceded by a backslash (\) in the definition.

For example, if you were trying to extract Web addresses, the full stop character is very important to the entity, therefore, you must backslash it such as:

```
www\.[a-z]+\.[a-z]+
```

Repetition Operators and Quantifiers ? + * {}

To enable the definitions to be more flexible, you can use several wildcards that are standard to regular expressions. They are * ? +

- **Asterisk *** indicates that there are *zero or more* of the preceding string. For example, `ab*c` matches “*ac*”, “*abc*”, “*abbc*”, and so on.
- **Plus sign +** indicates that there is *one or more* of the preceding string. For example, `ab+c` matches “*abc*”, “*abbc*”, “*abbbc*”, but not “*ac*”.
- **Question mark ?** indicates that there is *zero or one* of the preceding string. For example, `modell?ing` matches both “*modeling*” and “*modeling*”.
- **Limiting repetition with brackets {}** indicates the bounds of the repetition. For example,
 - ▶ `[0-9]{n}` matches a digit repeated exactly *n* times. For example, `[0-9]{4}` will match “*1998*”, but neither “*33*” nor “*19983*”.
 - ▶ `[0-9]{n,}` matches a digit repeated *n or more* times. For example, `[0-9]{3,}` will match “*199*” or “*1998*”, but not “*19*”.
 - ▶ `[0-9]{n,m}` matches a digit repeated between *n and m times inclusive*. For example, `[0-9]{3,5}` will match “*199*”, “*1998*” or “*19983*”, but not “*19*” nor “*199835*”.

Optional Spaces and Hyphens

In some cases, you want to include an optional space in a definition. For example, if you wanted to extract currencies such as “*uruguayan pesos*”, “*uruguayan peso*”, “*uruguay pesos*”, “*uruguay peso*”, “*pesos*” or “*peso*”, you would need to deal with the fact that there may be two words separated by a space. In this case, this definition should be written as `(uruguayan |uruguay)?pesos?`. Since *uruguayan* or *uruguay* are followed by a space when used with *pesos/peso*, the optional space must be defined within the optional sequence `(uruguayan |uruguay)`. If the space was not in the optional sequence such as `(uruguayan|uruguay)? pesos?`, it would not match on “*pesos*” or “*peso*” since the space would be required.

If you are looking for a series of things including a hyphen characters (-) in a list, then the hyphen must be defined last. For example, if you are looking for a comma (,) or a hyphen (-), use `[, -]` and never `[-,]`.

Order of Strings in Lists and Macros

You should always define the longest sequence before a shorter one or else the longest will never be read since the match will occur on the shorter one. For example, if you were looking for strings “*billion*” or “*bill*”, then “*billion*” must be defined before “*bill*”. So for instance `(billion|bill)` and not `(bill|billion)`. This also applies to macros, since macros are lists of strings.

Order of Rules in the Definition Section

Define one rule per line. Within each section, rules are numbered (*regexp1*, *regexp2*, and so on). These rules must be numbered sequentially from 1–*n*. Any break in numbering will cause the processing of this file to be suspended altogether. To disable an entry, place a # symbol at the beginning of each line used to define the regular expression. To enable an entry, remove the # character before that line.

In each section, the most specific rules must be defined before the most general ones to ensure proper processing. For example, if you were looking for a date in the form “*month year*” and in the form “*month*”, then the “*month year*” rule must be defined before the “*month*” rule. Here is how it should be defined:

```
#@# January 1932
regexp1=$(MONTH),? [0-9]{4}

#@# January
regexp2=$(MONTH)
```

and not

```
#@# January
regexp1=$(MONTH)

#@# January 1932
regexp2=$(MONTH),? [0-9]{4}
```

Using Macros in Rules

Whenever a specific sequence is used in several rules, you can use a macro. Then, if you need to change the definition of this sequence, you will need to change it only once, and not in all the rules referring to it. For example, assuming you had the following macro:

```
MONTH=(january|february|march|april|june|july|august|september|october|
november|december)|(jan|feb|mar|apr|may|jun|jul|aug|sep|oct|nov|dec)(\.)?
```

Whenever you refer to the name of the macro, it must be enclosed in \$(), such as:

```
regexp1=$(MONTH)
```

All macros must be defined in the [macros] section.

Normalization

When extracting nonlinguistic entities, the entities encountered are normalized to group like entities according to predefined formats. For example, currency symbols and their equivalent in words are treated as the same. The normalization entries are stored in the Normalization section in the Advanced Resources tab. For more information, see the topic “About Advanced Resources” on p. 225. The file is broken up into distinct sections.

Important! This file is for advanced users only. It is highly unlikely that you would need to change this file. If you require additional assistance in this area, please contact IBM Corp. for help.

Formatting Rules for Normalization

- Add only one normalization entry per line.

- Strictly respect the sections in this file. No new sections can be added.
- To disable an entry, place a # symbol at the beginning of that line. To enable an entry, remove the # character before that line.

Configuration

You can enable and disable the nonlinguistic entity types that you want to extract in the nonlinguistic entity configuration file. By disabling the entities that you do not need, you can decrease the processing time required. This is done in the Configuration section in the Advanced Resources tab. For more information, see the topic “About Advanced Resources” on p. 225. If nonlinguistic extraction is enabled, the extraction engine reads this configuration file during the extraction process to determine which nonlinguistic entity types should be extracted.

The syntax for this file is as follows:

```
#name<TAB>Language<TAB>Code
```

Table 11-1
Syntax for configuration file

Column label	Description
#name	The wording by which nonlinguistic entities will be referenced in the two other required files for nonlinguistic entity extraction. The names used here are case sensitive.
Language	The language of the records. It is best to select the specific language; however, an Any option exists. Possible options are: 0 = Any which is used whenever a regexp is not specific to a language and could be used in several templates with different languages, for instance an IP/URL/email addresses; 1 = French; 2 = English; 4 = German; 5 = Spanish; 6 = Dutch; 8 = Portuguese; 10 = Italian.
Code	Part-of-speech code. Most entities take a value of “s” except in a few cases. Possible values are: s = stopword; a = adjective; n = noun. If enabled, nonlinguistic entities are first extracted and the extraction patterns are applied to identify its role in a larger context. For example, percentages are given a value of “a.” Suppose that 30% is extracted as a nonlinguistic entity. It would be identified as an adjective. Then if your text contained “30% salary increase,” the “30%” nonlinguistic entity fits the part-of-speech pattern “ann” (adjective noun noun).

Order in Defining Entities

The order in which the entities are declared in this file is important and affects how they are extracted. They are applied in the order listed. Changing the order will change the results. The most specific nonlinguistic entities must be defined before more general ones.

For example, the nonlinguistic entity “Aminoacid” is defined by:

```
regexp1=($(AA)-?$(NUM))
```

where \$(AA) corresponds to

“(ala|arg|asn|asp|cys|gln|glu|gly|his|ile|leu|lys|met|phe|pro|ser)”, which are specific 3-letter sequences corresponding to particular amino acids.

On the other hand, the nonlinguistic entity “Gene” is more general and is defined by:

```
regexp1=p[0-9]{2,3}
regexp2=[a-z]{2,4}-?[0-9]{1,3}-?[r]
regexp3=[a-z]{2,4}-?[0-9]{1,3}-?p?
```

If “Gene” is defined before “Aminoacid” in the Configuration section, then “Aminoacid” will never be matched, since `regexp3` from “Gene” will always match first.

Formatting Rules for Configuration

- Use a TAB character to separate each entry in a column.
- Do not delete any lines.
- Respect the syntax shown in the preceding table.
- To disable an entry, place a # symbol at the beginning of that line. To enable an entity, remove the # character before that line.

Language Handling

Every language used today has special ways of expressing ideas, structuring sentences, and using abbreviations. In the Language Handling section, you can edit extraction patterns, force definitions for those patterns, and declare abbreviations for the language that you have selected in the Language drop-down list.

- Extraction patterns
- Forced definitions
- Abbreviations

Extraction Patterns

When extracting information from your records, the extraction engine applies a set of parts-of-speech extraction patterns to a “stack” of words in the text to identify candidate terms (words and phrases) for extraction. You can add or modify the extraction patterns.

Parts of speech include grammatical elements, such as nouns, adjectives, past participles, determiners, prepositions, coordinators, first names, initials, and particles. A series of these elements makes up a part-of-speech extraction pattern. In IBM Corp. text mining products, each part of speech is represented by a single character to make it easier to define your patterns. For instance, an adjective is represented by the lowercase letter *a*. The set of supported codes appears by default at the top of each default extraction patterns section along with a set of patterns and examples of each pattern to help you understand each code that is used.

Formatting Rules for Extraction Patterns

- One pattern per line.
- Use # at the beginning of a line to disable a pattern.

The order in which you list the extraction patterns is very important because a given sequence of words is read only once by the extraction engine and is assigned to the first extraction patterns for which the engine finds a match.

Forced Definitions

When extracting information from your records, the extraction engine scans the text and identifies the part of speech for every word it encounters. In some cases, a word could fit several different roles depending on the context. If you want to force a word to take a particular part-of-speech role or to exclude the word completely from processing, you can do so in the Forced Definition section of the Advanced Resources tab. For more information, see the topic “About Advanced Resources” on p. 225.

To force a part-of-speech role for a given word, you must add a line to this section using the following syntax:

```
term:code
```

Table 11-2
Syntax description

Entry	Description
<code>term</code>	A term name.
<code>code</code>	A single-character code representing the part-of-speech role. You can list up to six different part-of-speech codes per uniterm. Additionally, you can stop a word from being extracted into compound words/phrases by using the lowercase code <code>s</code> , such as <code>additional:s</code> .

Formatting Rules for Forced Definitions

- One line per word.
- Terms cannot contain a colon.
- Use the lowercase `s` as a part-of-speech code to stop a word from being extracted altogether.
- Use up to six part-of-speech codes per line. Supported part-of speech codes are shown in the Extraction Patterns section. For more information, see the topic “Extraction Patterns” on p. 233.
- Use the asterisk character (*) as a wildcard at the end of a string for partial matches. For example, if you enter `add*:s`, words such as `add`, `additional`, `additionally`, `addendum`, and `additive` are never extracted as a term or as part of a compound word term. However, if a word match is explicitly declared as a term in a compiled dictionary or in the forced definitions, it will still be extracted. For example, if you enter both `add*:s` and `addendum:n`, `addendum` will still be extracted if found in the text.

Abbreviations

When the extraction engine is processing text, it will generally consider any period it finds as an indication that a sentence has ended. This is typically correct; however, this handling of period characters does not apply when abbreviations are contained in the text.

If you extract terms from your text and find that certain abbreviations were mishandled, you should explicitly declare that abbreviation in this section.

Note: If the abbreviation already appears in a synonym definition or is defined as a term in a type dictionary, there is no need to add the abbreviation entry here.

Formatting Rules for Abbreviations

- Define one abbreviation per line.

Japanese Text Exceptions

While Japanese language text is processed and mined in a similar way to other supported languages in IBM® SPSS® Text Analytics for Surveys, there are a number differences. The smaller differences are described along side the instructions for all other languages in this documentation. However, some of the larger differences are covered in this appendix chapter.

Extracting and Categorizing Japanese Text

When mining Japanese language text, the process is similar to other supported languages. For more information, see the topic “About Text Mining” in Chapter 1 on p. 3. However, there are some differences for Japanese language as follows.

How Extraction Works

During the extraction of key concepts and ideas from your responses, IBM® SPSS® Text Analytics for Surveys relies on linguistics-based text analysis. This approach offers the speed and cost effectiveness of statistics-based systems. But it offers a far higher degree of accuracy, while requiring far less human intervention. Linguistics-based text analysis is founded on the field of study known as natural language processing, also known as computational linguistics.

For Japanese language text, the difference between statistics-based and linguistics-based approaches during the extraction process can be illustrated using the word 沈む as an example. Using this word, we can find expressions such as 日が沈む, translated as *sun goes down*, or 気分が沈む, translated as *feel down*. If you use statistical techniques alone, 日 (translated as *sun*), 気分 (translated as *feel*), and 沈む (translated as *down*) are each separately extracted. However, when we use the Sentiment analyzer, which uses linguistic techniques, not only are 日, 気分, and 沈む extracted, but 気分が沈む (translated as *feel down*) is extracted and assigned to the type <悪い - 悲しみ全般>. The use of linguistic-based techniques through the Sentiment analyzer make it possible to extract more meaningful expressions. The analysis and capture of emotions cuts through the ambiguity of text, and makes linguistics-based text mining, by definition, the more reliable approach.

Understanding how the extraction process works can help you make key decisions when fine-tuning your linguistic resources (libraries, types, synonyms, and more). Steps in the extraction process include:

- Converting source data to a standard format
- Identifying candidate terms
- Identifying equivalence classes and integration of synonyms
- Assigning a type
- Indexing and, when requested, pattern matching with a secondary analyzer

Step 1. Converting source data to a standard format

In this first step, the data you import is converted to a uniform format that can be used for further analysis. This conversion is performed internally and does not change your original data.

Step 2. Identifying candidate terms

It is important to understand the role of linguistic resources in the identification of candidate terms during linguistic extraction. Linguistic resources are used every time an extraction is run. They exist in the form of templates, libraries, and compiled resources. Libraries include lists of words, relationships, and other information used to specify or tune the extraction. The compiled resources cannot be viewed or edited. However, the remaining resources can be edited in the Resource Editor.

Compiled resources are core, internal components of the extraction engine within SPSS Text Analytics for Surveys. These resources include a general dictionary containing a list of base forms with a part-of-speech code (noun, verb, adjective, and so on). The resources also include reserved, built-in types used to assign many extracted terms to the following types, <地名>, <組織>, or <人名>. For more information, see the topic “Available Types for Japanese Text” on p. 246.

In addition to those compiled resources, several libraries are delivered with the product and can be used to complement the types and concept definitions in the compiled resources, as well as to offer synonyms. These libraries—and any custom ones you create—are made up of several dictionaries. These include type dictionaries, synonym dictionaries, and exclude dictionaries. For more information, see the topic “Editing Resources for Japanese Text” on p. 242.

Once the data have been imported and converted, the extraction engine will begin identifying candidate terms for extraction. Candidate terms are words or groups of words that are used to identify concepts in the text. During the processing of the text, single words (**uniterms**) and compound words (**multiterms**) are identified using part-of-speech pattern extractors. For example, the multiterm 青森りんご, which follows the <地名> + <名詞> part-of-speech pattern, has two components. Then, candidate sentiment keywords are identified using sentiment text link analysis.

For example, let’s say you have the following text in Japanese: 写真が新鮮で良かった。In this case, the extraction engine would assign the sentiment type 良い - 褒め・賞賛, after matching (品物) + が + 良い using one of the sentiment text link rules.

Note: The terms in the aforementioned compiled general dictionary represent a list of all of the words that are likely to be uninteresting or linguistically ambiguous as uniterms. These words are excluded from extraction when you are identifying the uniterms. However, they are reevaluated when you are determining parts of speech or looking at longer candidate compound words (multiterms).

Step 3. Identifying equivalence classes and integration of synonyms

After candidate uniterms and multiterms are identified, the software uses a normalization dictionary to identify equivalence classes. An equivalence class is a base form of a phrase or a single form of two variants of the same phrase. The purpose of assigning phrases to equivalence classes is to ensure that, for example, *side effect* and 副作用 are not treated as separate concepts. To determine which concept to use for the equivalence class—that is, whether *side effect* or 副作用 is used as the lead term—the extraction engine applies the following rules in the order listed:

- The user-specified form in a library.
- The most frequent form, as defined by precompiled resources.

Step 4. Assigning type

Next, types are assigned to extracted concepts. A type is a semantic grouping of concepts. Both compiled resources and the libraries are used in this step. Types include such things as higher-level concepts, positive and negative words, first names, places, organizations, and more. For more information, see the topic “Type Dictionaries” in Chapter 10 on p. 207.

Japanese language resources have a distinct set of types. For more information, see the topic “Available Types for Japanese Text” on p. 246.

Step 5. Indexing and pattern matching with event extraction

The entire set of records is indexed by establishing a pointer between a text position and the representative term for each equivalence class. This assumes that all of the inflected form instances of a candidate concept are indexed as a candidate base form. The global frequency is calculated for each base form.

SPSS Text Analytics for Surveys can discover not only types and concepts but also relationships among them. Several algorithms and libraries are available with this product and provide the ability to extract text link analysis relationship patterns between types and concepts. They are particularly useful when attempting to discover specific opinions (for example, product reactions).

How Secondary Extraction Works

When you perform an extraction on Japanese text, you automatically obtain concepts from the basic keywords and 8 basic types, including 人名, 地名, 組織名, 名詞, 形容詞, 動詞, 形容動詞, and その他. However, in order to fully benefit from the default resources provided for Japanese text, you must select one of the following secondary analyzers: Sentiment or Dependency.

Choosing a secondary analyzer also enables you to extract text link analysis patterns and uncover relationships between the terms in the text.

Secondary Analysis. When an extraction is launched, basic keyword extraction takes place using the default set of types. For more information, see the topic “Available Types for Japanese Text” on p. 246. However, when you select a secondary analyzer, you can obtain many more or richer concepts since the extractor will now include particles and auxiliary verbs as part of the concept. For example, let’s assume we had the sentence 肩の荷が下りた, translated as “*I have a great weight lifted from my shoulders*”. With this example, the basic keyword extraction can extract each concept separately such as: 肩 (*shoulders*), 荷 (*weight*), 下りる (*have lifted*), but the relationship between these words is not extracted. However, if you applied Sentiment analysis, you can extract richer concepts relating to a sentiment type such as the concept =肩の荷が下りた, which is translated as “*have a great weight lifted from one’s shoulders*”, assigned to the type <良い-安心>. In the case of sentiment analysis, a large number of additional types are also included. Furthermore, choosing a secondary analyzer allows you to also generate text link analysis results.

Note: When a secondary analyzer is called, the extraction process takes longer to complete. For more information, see the topic “How Secondary Extraction Works” on p. 239.

- **Dependency analysis.** Choosing this option yields extended particles for the extraction concepts from the basic type and keyword extraction. You can also obtain the richer pattern results from dependency text link analysis (TLA).
- **Sentiment analysis.** Choosing this analyzer yields additional extracted concepts and, whenever applicable, the extraction of TLA pattern results. In addition to the basic types, you also benefit from more than 80 sentiment types, including 嬉しい, 吉報, 幸運, 安心, 幸福, and so on. These types are used to uncover concepts and patterns in the text through the expression of emotion, sentiments, and opinions. There are three options that dictate the focus for the sentiment analysis: All sentiments, Representative sentiment only, and Conclusions only.

Sentiment Analysis Options

When working with Japanese text, you can choose to extract additional concepts and types using the Sentiment analyzer. This analyzer includes over 80 additional types to help you extract opinions, feelings, and emotions from your text data. Additionally, when you choose Sentiment analysis as your secondary analyzer, you must also select one of the following options, which tell the extraction engine which sentiments to extract:

- **All sentiments**
- **Representative sentiment only**
- **Conclusions only**

During extraction, the sentiment analyzer begins by dividing a record into clauses, each of which contains a predicate. For example, the text, “4月になったがまだ寒い。”, which is translated as “*It’s April, but it’s still cold.*”, is interpreted as 2 clauses by the analyzer despite the fact that it contains only one stop character 。. Each clause is then examined by the extraction engine to see if it fits the option you selected.

Let’s examine the three options using the sample text: “案内してくれた仲居さんは無愛想だったが、部屋は広くて申し分なかった。夕食も満足。”. This text is translated as: “*A serving lady was not friendly, but the room was large and quite satisfactory. I was satisfied with the dinner, too.*”. During extraction, the original text is broken into the following clauses:

- 案内してくれた仲居さんは無愛想だったが、, which means “*A serving lady was not friendly, but*”
- 部屋は広くて申し分なかった。 , which means “*The room was large and quite satisfactory*”
- 夕食も満足。 , which means “*I satisfied with the dinner, too.*”

All Sentiments

This option extracts all sentiments, opinions, and emotions that match the resources and sentiment text link rules. With our sample, the following concepts could be extracted from the sample text.

Table A-1
Possible output for the sample using the “All Sentiments” option

Concept	Type
仲居さんは無愛想だった	<悪い - 対応が不親切>
部屋は広くて	<良い - 満足>

Concept	Type
申し分なかった	<良い - 満足>
満足	<良い - 満足>

Note: In the preceding table, the second and third rows show how the extractor might obtain two concepts from the same clause.

Representative Sentiment Only

This option extracts only the more representative opinions or emotions expressed in each clause. If there are several opinions or emotions in the text, an algorithm is applied. This algorithm attempts to determine the importance of the sentiments found, and the position of the words in a clause. In some cases where two sentiment keywords with the same importance are found, the sentiment keyword in the last position in the clause is extracted rather than the first.

部屋は広くて, which is translated as *the room was wide*, is not extracted from the text since the second *申し分なかった* is considered more important than the first *部屋は広くて* in this clause when the internal algorithm and word position is applied.

Table A-2

Possible output for the text using the "Representative Sentiment Only" option

Concept	Type
仲居さんは無愛想だった	<悪い - 対応が不親切>
申し分なかった	<満足>
満足	<満足>

Conclusions Only

This option forces the extractor to identify and extract a sentiment keyword as representing the conclusion of the entire record. Not all text has a conclusion so in some cases, nothing may be extracted for a given piece of text with this option. Additionally, the longer the record, the harder it is for the analyzer to identify the main conclusion. While rare, it is still possible for multiple conclusions to be extracted.

満足, which is translated as *satisfied*, is considered to be the essential conclusion of the sentiments expressed in the text.

Table A-3

Possible output for the text using the "Conclusions Only" option

Concept	Type
満足	<満足>

How Categorization Works

There are several different techniques you can choose to create categories. Because every dataset is unique, the number of techniques and the order in which you apply them may change. Since your interpretation of the results may be different from someone else's, you may need to experiment with the different techniques to see which one produces the best results for your text data.

In this guide, **category building** refers to the generation of category definitions and classification through the use of one or more built-in techniques, and **categorization** refers to the scoring, or labeling, process whereby unique identifiers (name/ID/value) are assigned to the category definitions for each record.

During category building, the concepts and types that were extracted are used as the building blocks for your categories. When you build categories, the records are automatically assigned to categories if they contain text that matches an element of a category's definition.

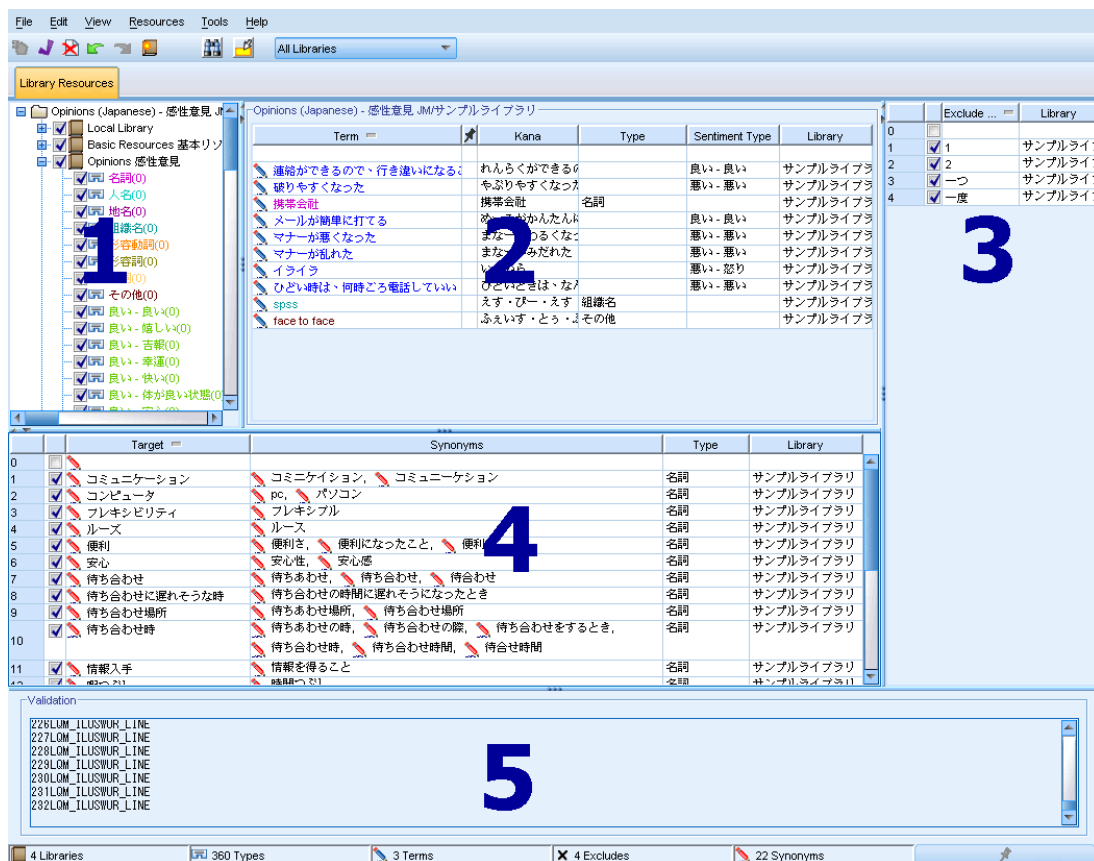
IBM® SPSS® Text Analytics for Surveys offers you several automated category building techniques to help you categorize your records quickly. Each of the techniques is well-suited to certain types of data and situations, but often it is helpful to combine techniques in the same analysis to capture the full range of records. You may see a concept in multiple categories or find redundant categories.

Editing Resources for Japanese Text

Beginning in IBM® SPSS® Text Analytics for Surveys version 4, a new template and text analysis package (TAP) are available for Japanese text. You can make changes to the resources by adding and editing terms to customize them to your data. The text analysis package also contains a category set made up of categories representing positive sentiments, negative sentiments, and contextual/generic sentiments.

You can work with your resources in the Resource Editor. Editors work similarly for all text languages; however, there are some significant differences for Japanese text as described here.

Figure A-1
Resource Editor view for Japanese Text



The following points highlight some of the key differences when working with resources for Japanese text. For a general description of the four main panes in the Library Resources tab, see “The Editor Interface” on p. 184.

- 1. Library pane.** Located in the upper left corner, this area works much like it does for other languages. However, there are a few differences such as not being able to create new types or rename types. For more information, see the topic “Working with Libraries” in Chapter 9 on p. 195.
- 2. Term pane for type dictionaries.** Located to the right of the library tree pane, this pane is quite different for Japanese text. In addition to having the term name, you can also add the Kana name as well as select one or two types to which you can associate the term. However, you cannot generate inflected forms of terms or assign match options for Japanese terms like you can for non-Japanese languages. For more information, see the topic “Japanese Library Tree, Types, and Term Pane” on p. 244.
- 3. Substitution/Synonym dictionary pane.** In Japanese text resources, you will find one Synonym tab in which you can define all the synonyms for your resources. In the Synonym tab, there is an additional Type column in which you must designate a type for the synonyms entered. For more

information, see the topic “Using the Synonym Dictionary for Japanese Text” on p. 250. *Note:* The Optional Elements tab does not appear because it does not apply to Japanese text.

4. Exclude dictionary pane. There are no differences in this pane for Japanese text resources except that the use of the * wildcard is not supported.

5. Validation pane. For Japanese text, there is an additional validation pane used to check your resources before extraction. When extracting from Japanese text, the extraction engine automatically recompiles the resources if changes are detected before beginning the extraction process. To avoid potential errors during extraction, you can recompile and validate the resources before extracting so that you can correct any errors encountered. For more information, see the topic “Validating and Compiling Japanese Resources” on p. 252.

Note: There are no advanced resources or text link rules that are editable for Japanese language text so these tabs are not available.

Japanese Library Tree, Types, and Term Pane

The way you work with libraries and types for Japanese resources is much like it is for other languages. For more information, see the topic “Type Dictionaries” in Chapter 10 on p. 207.

However, there are a few main differences, including:

- Japanese text resources have a distinct set of types. For more information, see the topic “Available Types for Japanese Text” on p. 246.
- Types cannot be created or renamed; however their properties can be edited. For more information, see the topic “Editing Japanese Type Properties” on p. 250.
- You can add and edit terms, including the specification of a Kana name for a term as well as the assignment to a type and a secondary sentiment type. For more information, see the topic “Japanese Library Tree, Types, and Term Pane” on p. 244.

The library tree pane displays the libraries as well as their type dictionaries. If you select a library or type in the left hand pane, a term pane to the right displays the terms for the selected libraries or type dictionaries. You can add terms to a type dictionary directly in the term pane or through the Add Terms dialog box. The terms that you add can be single words or compound words. You will always find a blank row at the top of the list to allow you to add a new term.

When you define a term in a type dictionary, it is considered to be a noun by default and automatically assigned to the type <名詞>. However, you can change the type to another basic type such as <動詞>, <形容詞>, <地名> and so on. If the extraction engine finds this term acting as the same part-of-speech as the type to which you assigned it in the Type column, then it will be assigned to that type and extracted. You can additionally assign the term to one of the sentiment types in the Sentiment Type column. Then, when you use the Sentiment secondary analyzer, then the text is processed a second time in order to try to find terms and assign them to the sentiment types. Furthermore, if you define both a sentiment type and basic type, and the extraction engine finds this term matching both types when secondary sentiment analysis was also performed, then the sentiment type takes precedence and is shown in the extraction results pane and text link analysis results. For example, if a verb was extracted as a verb <動詞> type and also as a positive

kind of type such “loved”, then this term would be shown as belonging to the positive type in the interface since capturing sentiments is oftentimes more interesting than just a part-of-speech.

Figure A-2
Library and Term panes for Japanese resources

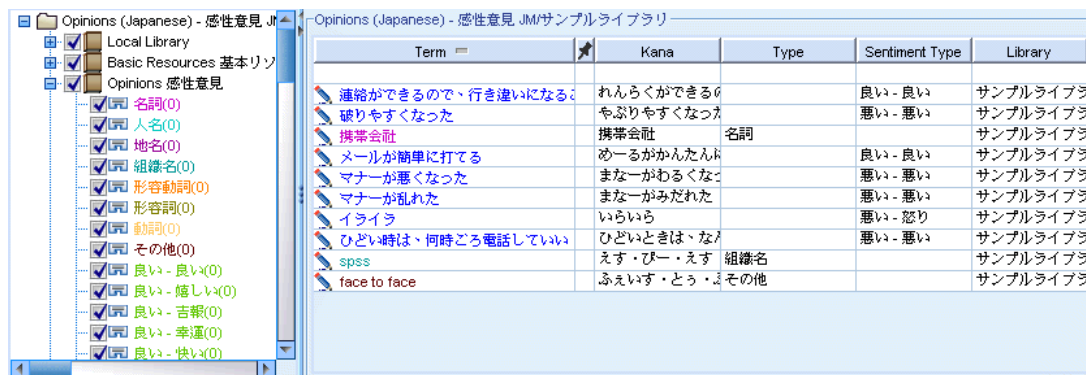


Table A-4
Term Pane Column Descriptions

Column Name	Column Description
Term	Enter single or compound words into the cell. The color in which the term appears depends on the color for the type in which the term is stored or forced. You can change type colors in the Type Properties dialog box. For more information, see the topic “Editing Japanese Type Properties” on p. 250. Generally, the term is written in Kanji but may also comprise Kana. Important! Entering verbs using Katakana characters is not supported.
Force	Clicking and placing a pushpin icon into this cell lets the extraction engine know to ignore any other occurrences of this same term in other libraries. For more information, see the topic “Forcing Terms” in Chapter 10 on p. 214. This works the same for all languages.
Kana	Enter the Kana spelling of the Kanji term name.
Type	Select the basic type name to which the term should be assigned. For more information, see the topic “Available Types for Japanese Text” on p. 246.
Sentiment Type	If secondary analysis will be performed, select the sentiment type name to which the term should be assigned. For more information, see the topic “Available Types for Japanese Text” on p. 246.
Library	Select the library in which your term is stored. You can drag and drop a term into another type in the library tree pane to change its library.

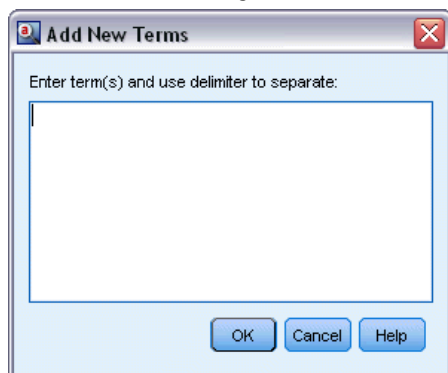
To Add a Single Term to a Type Dictionary

- ▶ In the library tree pane, select the type dictionary to which you want to add the term.
- ▶ In the term list in the center pane, type your term in the first available empty cell and set any options you want for this term.

To Add Multiple Terms to a Type Dictionary

- ▶ In the library tree pane, select the type dictionary to which you want to add terms.
- ▶ From the menus, choose Tools > New Terms. The Add New Terms dialog box opens.

Figure A-3
Add New Terms dialog box



- ▶ Enter the terms you want to add to the selected type dictionary by typing the terms or pasting a set of terms. If you enter multiple terms, you must separate them using the delimiter that is defined in the Options dialog, or add each term on a new line. For more information, see the topic “Setting Options” in Chapter 2 on p. 16.
- ▶ Click OK to add the terms to the dictionary. The dialog box closes and the new terms appear in the dictionary.

Available Types for Japanese Text

You cannot add new types to the Japanese resources, however you can add and remove terms from them. The following tables includes the set of Japanese types currently available.

Types for Basic Extraction

Whenever an extraction is launched, the following types are used.

Table A-5
Types for Basic Extraction

Types	Description
名詞	Words that refer to things, such as “car” and “movie.” Personal names, place names, and organizational names, however, are categorized separately.
人名	Nouns that correspond to the names of specific people, such as “Tokugawa” and “Ieyasu.” Combinations of first and last names, like “Tokugawa Ieyasu,” are also personal names.
地名	Nouns such as “Tokyo” and “London” that refer to specific places.
組織名	Nouns that refer to particular companies and organizations, such as “The Federation of Economic Organizations.”
形容動詞	Words like “quiet (shizuka)” that describe the characteristics or condition of a thing and can be used in “not [adjective] (~ de nai)” and “a [adjective] thing (~ na koto)” phrases.
形容詞	Words like “fun (tanoshii)” that describe the characteristics or condition of a thing and can be used in phrases such as “become [adjective] (~ku naru)” and “a [adjective] thing (~i koto).”

Types	Description
動詞	Words that describe movement or action, including type I (consonant stem) verbs, type II (vowel stem) verbs, and irregular (sagyou henkaku and kagyou henkaku) verbs.
その他	Words such as adverbs, pre-noun adjectivals, conjunctions, and interjections; examples include “quite,” “whatsoever,” “then,” and “thank you.”

Types for Sentiment Analysis

Whenever you select the secondary analyzer for sentiment extraction, you get a large number of types in addition to the 8 basic types.

Table A-6
Types for Sentiment Analysis

Types	Description
良い - 良い	Expressions of generally positive things that can be classified as “good.”
良い - 嬉しい	Describes a desirable event that produces a pleasant stimulation.
良い - 吉報	Describes a pleasant event that can only be made possible by considerable effort.
良い - 幸運	Describes a happy event that can only be made possible by chance or a remarkable coincidence.
良い - 快い	Suggests that something is a stimulus or environment that triggers a pleasant physiological sensation.
良い - 体が良い状態	Describes a state in which the body is free of sickness, injury, and fatigue, or a state in which physical condition is improving.
良い - 安心	Suggests that one is calm and at no risk of harm or damage.
良い - 幸福	Indicates that one has obtained especially favorable conditions or affection through one’s own action or the circumstances of one’s birth.
良い - 満足	Describes a desirable event that calms the mind.
良い - 美味しい	Indicates that a food has a pleasant taste.
良い - 効果が満足	Implies that a certain thing has produced the expected effect.
良い - 感動	Suggests that the significance, meaning or value of something is astonishingly good.
良い - 感謝	Suggests that one recognizes the actions of another in a positive way.
良い - 祝福	Expresses the view that another person’s situation is favorable (to a degree acceptable to the speaker).
良い - 喜び全般	Otherwise positive events or positive events with little connection to the speaker.
良い - 楽しい	Indicates or anticipates activities such as companionship, amusement, and recreation.
良い - 可笑しい	Denotes that something has a humorous quality that provides a pleasant stimulation.
良い - 笑い	Expresses a smile or laughter caused by a good and/or humorous thing.
良い - 期待	Predicts that a good event will occur in the future.
良い - 楽しみ全般	Otherwise enjoyable events and/or positive activities/behavior with little connection to the speaker.
良い - 金額への賞賛	Implies that, from the buyer’s standpoint, something has a desirable monetary value.
良い - 対応が早い	Suggests that a service was provided or completed in a timely manner.
良い - 対応が親切	Suggests that the attitude or behavior of the provider of a service was solicitous.

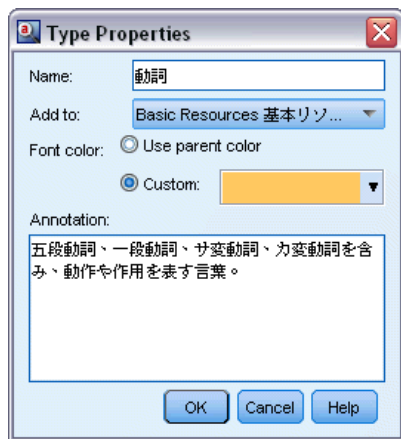
Types	Description
良い - 説明が良い	Expresses the idea that the type and/or quantity of information, and/or the method of its provision, is appropriate.
良い - 対応への賞賛	Views, other than those above, that praise the provider of a service.
良い - 褒め・賞賛	Views, other than those above, that praise the characteristics, capabilities, and/or operation of a certain thing.
良い - 好き	Expresses the desire to possess or grow close to a certain thing.
良い - 入会希望	Describes the desire to be or remain part of a certain group.
良い - 買いたい	Implies that one wants to or plans to use money to obtain a certain thing.
良い - 好評・人気	Indicates that the number of people who want or appreciate a certain thing has exceeded a certain goal.
良い - 売れた	Indicates the presence of people who purchase a certain thing or that the number or value of purchases has exceeded a certain goal.
悪い - 悪い	Expressions of generally negative things that can be classified as “bad.”
悪い - 怒り	A distinct sense of anger felt when something does not happen as one had planned.
悪い - 批判	Expresses the idea that another person has failed to make the appropriate choice.
悪い - お叱り	Words or actions that intimidate another person to conform to one’s intentions.
悪い - 誹謗・中傷	Words used to demonstrate an excessively low opinion of another person.
悪い - 軽蔑	Implies that another person’s character, abilities, and/or other qualities is/are severely lacking.
悪い - 恨み	Expresses retribution or resentment for a disadvantage caused by another person.
悪い - 嫌がらせ	Words used for the purpose of inhibiting communication.
悪い - 不満	An unpleasant feeling caused by the inability to obtain the desired thing or state.
悪い - 不味い	Indicates that a food has a bad taste.
悪い - 効果が不満	Implies that something has not produced the expected effect.
悪い - 金額が不満	Implies that, from the buyer’s standpoint, the monetary value of a certain thing is undesirable.
悪い - 対応への不満	Suggests that the provider of a service is at fault.
悪い - 対応が遅い	Implies that a service has been performed / completed in an untimely manner, or that the service has yet to be completed.
悪い - 対応が不親切	Denotes an unpleasant feeling caused by the attitude or behavior of the provider of a service.
悪い - 説明が悪い	Expresses the idea that the type and/or quantity of information, and/or the method of its provision, is inappropriate.
悪い - 返答なし	Denotes that the provider of a service fails to provide a proper response, even though the situation demands one.
悪い - 不快	Suggests that something is a stimulus or environment that triggers a negative physiological sensation.
悪い - 怒り全般	Feelings of anger other than those above. General anger experienced by the speaker’s organization or company, or descriptions of events caused by said anger.
悪い - 悲しい	A distinct unpleasant feeling experienced when one loses or cannot obtain something.
悪い - 凶報	Expresses the idea that a certain goal could not be achieved despite substantial effort.
悪い - 不運	Indicates a negative outcome caused by unfortunate coincidence and/or luck, not one’s own fault.
悪い - ショック	Suggests that one is upset by an unforeseen, negative thing or occurrence and is unable to find an appropriate response.

Types	Description
悪い - 残念	An unhappy feeling experienced when something that is anticipated to happen fails to do so.
悪い - 落胆	A state in which one is overcome by an unhappy, disappointed feeling.
悪い - 諦め	Suggests that a negative thing, experienced by either the speaker or another person, cannot be improved.
悪い - 後悔	Expresses the idea that in the past, one failed to make the appropriate choice, even though it was available as an option.
悪い - 謝罪	Indicates the speaker's recognition of having caused harm to another.
悪い - 淋しい	Expresses the idea that contact with others is scarce or that others with whom contact can be made are few in number.
悪い - 哀れみ	Expresses the idea that another's situation is significantly worse than the speaker's.
悪い - 悩み	Indicates that one must make a choice but is unable to choose from the available options.
悪い - 困っている	Expresses the idea that there is no effective way to respond to a situation that demands action.
悪い - 苦しい	Expresses an unpleasant psychological state in which one is unable to act normally due to external causes or one's own error(s) or mistake(s).
悪い - 体が悪い状態	Describes a state in which the body is sick, injured, and/or fatigued, or a state in which physical condition is not improving.
悪い - 不安	Expresses the idea that something may not continue in its desirable state or may not satisfy expectations.
悪い - 恐怖	Suggests that a certain thing seems likely to cause one harm or injury.
悪い - 悲しみ全般	Feelings of sadness other than those above, such as general sadness about an unspecified thing.
悪い - 嫌い	Indicates that one wants to keep something away or move away from something.
悪い - 退会希望	Describes the desire to leave or refrain from joining a certain group.
悪い - 買いたくない	Suggests that one does not want a certain thing or does not plan to pay for said thing.
悪い - 不評・不人気	Indicates that the number of people who like a certain thing has failed to meet a certain goal or that there are many people who have negative feelings towards said thing.
悪い - 売れていない	Indicates the absence of people who purchase a certain thing or that the number or value of purchases has failed to meet a certain goal.
その他 - 疑問	Expressions that demand information requiring the other person's further examination or thought.
その他 - 問い合わせ	Expressions that demand information already in the other person's possession.
その他 - 要望	Expressions that command the other person (when the other person is at direct fault or of a lower rank than the speaker) to resolve a problem.
その他 - 提案・忠告	Expressions that command the other person (when the other person is at direct fault or of a lower rank than the speaker) to behave in a better way.
その他 - お願い	Expressions that command the other person, when the other person is not at fault or not of a lower rank than the speaker, to do something.
その他 - 激励	Expressions that encourage another person, or descriptions of encouraging behavior.
その他 - 勧誘	Expressions that command another person to do something together with the speaker.
その他 - 驚き	Expresses the idea that an event's suddenness or scale transcends rational judgment/understanding.
評価なし - 評価なし	No expression of evaluation.

Editing Japanese Type Properties

While you cannot create types in Japanese resources, you can view and edit type properties. Please note that the options such as the match option and inflected forms do not apply to Japanese text.

Figure A-4
Type Properties dialog box for Japanese text resources



Name. The name of the type dictionary.

Add to. This field indicates the library in which you will create your new type dictionary.

Font color. This field allows you to distinguish the results from this type from others in the interface. If you select Use parent color, the default type color is used for this type dictionary, as well. This default color is set in the options dialog box. For more information, see the topic “Options: Display Tab” in Chapter 2 on p. 18. If you select Custom, select a color from the drop-down list.

Annotation. This field is optional and can be used for any comments or descriptions.

To View or Edit Type Properties

- ▶ Select the type whose properties you want to see.
- ▶ Right-click your mouse and choose Type Properties from the context menu. The Type Properties dialog box opens.
- ▶ Make any necessary changes.
- ▶ Click OK to save the changes to the type dictionary.

Using the Synonym Dictionary for Japanese Text

For Japanese text, the substitution dictionary only contains one tab to manage your synonyms, the Synonym tab. Synonyms associate two or more words that have the same meaning. You can also use synonyms to group terms with their abbreviations or to group commonly misspelled words with the correct spelling.

Figure A-5
Synonym entries for Japanese text

	Target	Synonyms	Type	Library
0				
1	☑ コミュニケーション	☑ コミュニケーション, ☑ コミュニケーション	名詞	Local Library
2	☑ コンピュータ	☑ pc, ☑ パソコン	名詞	Local Library
3	☑ フレキシビリティ	☑ フレキシブル	名詞	Local Library
4	☑ ルーズ	☑ ルーズ	名詞	Local Library
5	☑ 便利	☑ 便利さ, ☑ 便利になったこと, ☑ 便利性	名詞	Local Library
6	☑ 安心	☑ 安心性, ☑ 安心感	名詞	Local Library
7	☑ 待ち合わせ	☑ 待ちあわせ, ☑ 待ち合わせ, ☑ 待ち合わせ	名詞	Local Library
8	☑ 待ち合わせに遅れそうな時	☑ 待ち合わせの時間に遅れそうなとき	名詞	Local Library
9	☑ 待ち合わせ場所	☑ 待ちあわせ場所, ☑ 待ち合わせ場所	名詞	Local Library
10	☑ 待ち合わせ時	☑ 待ちあわせの時, ☑ 待ち合わせの際, ☑ 待ち合わせをするとき, ☑ 待ち合わせ時, ☑ 待ち合わせ時間, ☑ 待ち合わせ時間	名詞	Local Library
11	☑ 情報入手	☑ 情報を得ること	名詞	Local Library

A synonym definition is made up of two parts. The target term is the term under which you want the extraction engine to group all synonym terms. Unless this target term is used as a synonym of another target term or it is excluded, it is likely to become the concept that appears in the Extraction Results pane. The list of synonyms are the terms that will be grouped under the target term.

On the Synonyms tab, you can enter a synonym definition in the empty line at the top of the table. Begin by defining the target term and its synonyms. You can also select the library in which you would like to store this definition. During extraction, all occurrences of the synonyms will be grouped under the target term in the final extraction. For more information, see the topic “Adding Terms” in Chapter 10 on p. 210.

When you are building your type dictionaries, you may enter a term and also have three or four synonyms in mind for that term. In that case, you could enter all of the terms and then your target term into the substitution dictionary and then drag the synonyms.

Important! Wildcards and special characters are not supported for Japanese text synonyms.

To Add a Synonym Entry

- ▶ In the empty line at the top of the table in Synonym tab of the Substitution pane, enter your target term in the Target column. The target term you entered appears in color. This color represents the type in which the term appears or is forced, if that is the case. If the term appears in black, this means that it does not appear in any type dictionaries.
- ▶ Click in the second cell to the right of the target and enter the set of synonyms. Separate each entry using the global delimiter as defined in the Options dialog box. All synonyms entered should be from the same type. For more information, see the topic “Setting Options” in Chapter 2 on p. 16. The terms that you enter appear in color. This color represents the type in which the term appears. If the term appears in black, this means that it does not appear in any type dictionaries.
- ▶ In the third column, the Type column, designate a type for these synonyms. The target, however, takes the type assigned during extraction. However, if the target was not extracted as a concept, then the type listed in this column is assigned to the target in the extraction results.
- ▶ Click in the last cell to select the library in which you want to store this synonym definition.

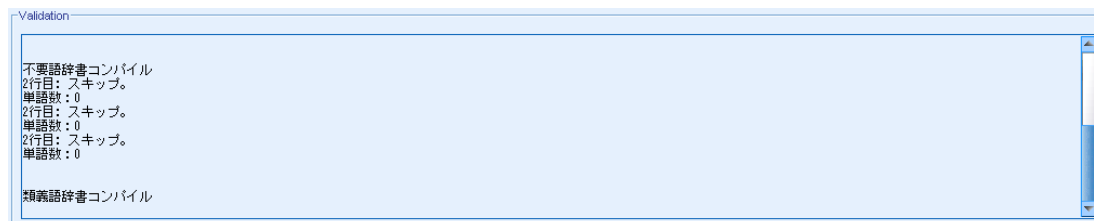
Note: These instructions show you how to make changes within the Resource Editor view. Keep in mind that you can also do this kind of fine-tuning directly from the Extraction Results pane or Data pane. For more information, see the topic “Refining Extraction Results” in Chapter 5 on p. 84.

Validating and Compiling Japanese Resources

For Japanese text, there is an additional validation pane used to check your resources before extraction. Before the extraction process begins for Japanese text, the extraction engine automatically recompiles the resources when changes are detected before beginning the extraction process. If an error is found during extraction, the process may not be able to complete correctly.

To avoid compilation errors, we recommend that you validate and compile your resources after you make changes in the Resource Editor. If any error messages appear, you can make corrections and attempt to validate again.

Figure A-6
Validation pane for Japanese text



To Validate Resources

- From the menus, choose Tools > Validate Resources. The validation pane opens to display compilation and error messages .

Other Exceptions for Japanese

Internal Resources Overriding User Defined Resources

For Japanese text, the default resources include some precompiled, internal basic resources. These internal resources are non-editable. For this reason, you can use the Resource Editor to make some changes and refinements. In almost all cases, any terms, synonyms, and exclude list entries you define in your resources will take precedence over the precompiled internal resources. However there are several exceptions as noted in some of the following examples.

- There are instances where adding terms to a particular type has no effect on the extraction results. This is most likely to occur when the data contains long sentences that include several morphological elements, punctuations or symbols. Additionally, since Japanese text resources already contain a large number of precompiled, common terms, there are some common words that will always be forced into a specific linguistic definition.
- You may be unable to exclude terms such as ある, いる, or なる since the extraction engine will always force the extraction of these terms.
- While it is possible to change the type of the term 東京 from <地名> to <名詞>, the extraction engine will ignore your change if you try to change the type of a term from <地名> to <動詞> or to <形容詞> using the Keyword (Type) dictionary.
- There may be times when changes you make in the Resource Editor or affect the extraction results from one sentence but not another sentence since the extraction process ends by referencing the co-occurrence words in each sentence.

Half-Width Katakana Display Issue

Half-width Katakana characters are internally converted to full-width Katakana characters during extraction but still appear in the original half-width Katakana characters when shown in the Data pane found in the user interface. Please note that half-width Katakana characters cannot be highlighted in the Data pane. To avoid this issue, convert your whole records to full-width Katakana before processing.

Upper and Lower Case Character Usage

Uppercase alphabetic characters are temporarily converted to lowercase alphabetic characters when read into the application. However, the Data pane will display the text using the same case as in the original text. Lowercase and uppercase characters are treated as the same in this product.

Notices

This information was developed for products and services offered worldwide.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785, U.S.A.

For license inquiries regarding double-byte character set (DBCS) information, contact the IBM Intellectual Property Department in your country or send inquiries, in writing, to:

Intellectual Property Licensing, Legal and Intellectual Property Law, IBM Japan Ltd., 1623-14, Shimotsuruma, Yamato-shi, Kanagawa 242-8502 Japan.

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Licenseses of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact:

IBM Software Group, Attention: Licensing, 233 S. Wacker Dr., Chicago, IL 60606, USA.

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The licensed program described in this document and all licensed material available for it are provided by IBM under terms of the IBM Customer Agreement, IBM International Program License Agreement or any equivalent agreement between us.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

If you are viewing this information softcopy, the photographs and color illustrations may not appear.

Trademarks

IBM, the IBM logo, [ibm.com](http://www.ibm.com), and SPSS are trademarks of IBM Corporation, registered in many jurisdictions worldwide. A current list of IBM trademarks is available on the Web at <http://www.ibm.com/legal/copytrade.shtml>.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Java and all Java-based trademarks and logos are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.

Microsoft product screenshot(s) reprinted with permission from Microsoft Corporation.

Other product and service names might be trademarks of IBM or other companies.

- ! ^ * \$ symbols in synonyms, 219
- & | !() rule operators, 147

- abbreviations, 233–234
- accommodating
 - punctuation errors, 82
 - spelling errors, 82
- activating nonlinguistic entities, 232
- adding
 - concepts to categories, 150
 - descriptors, 101
 - optional elements, 220
 - public libraries, 197
 - sounds, 18, 20
 - synonyms, 85, 218
 - synonyms for Japanese, 250
 - terms to exclude list, 223
 - terms to Japanese type dictionaries, 244
 - terms to type dictionaries, 210
 - types, 87
- addresses (nonlinguistic entity), 228
- advanced resources, 225
 - find and replace in editor, 226
- all documents, 93
- amino acids (nonlinguistic entity), 228
- AND rule operator, 147
- annotations
 - for categories, 104, 149
 - for projects, 48
- antilinks, 113
- assigning flags, 73
- asterisk (*)
 - exclude dictionary , 222
 - synonyms, 219
- auto- settings, 164
- autosaving projects, 17

- backing up resources, 190
- bar charts, 159
- blank responses, 26
- Boolean operators, 147
- Budget library, 208
- Budget type dictionary, 208
- building categories, 6, 105–106, 108–120, 122, 124, 242
 - peer/sibling grouping techniques, 122

- caret symbol (^), 219
- categories, 25, 91–92, 103, 148
 - adding to, 150
 - annotations, 104, 149
 - building, 6, 105, 109, 111, 113, 122
 - commonality charts, 159
 - copying, 156
 - creating, 98, 118, 125
 - creating new empty category, 124
 - deleting, 157
 - descriptors, 100–101, 104
 - editing, 148, 150
 - exporting, 52–53
 - extending, 113, 120
 - forcing responses, 153
 - forcing words, 154
 - labels, 104, 149
 - manual creation, 124
 - merging, 152
 - moving, 151
 - names, 104, 149
 - printing, 156
 - properties, 104, 149
 - refining results, 8, 148
 - relevance, 96–97
 - renaming, 124
 - scoring, 94
 - strategies, 99
 - text analysis packages, 40, 42
 - web graph, 159
- categories pane, 92
- categorizing, 6, 91, 242
 - co-occurrence rules, 109, 113, 117
 - concept inclusion, 109, 113, 115
 - concept root derivation, 109, 113–114
 - frequency techniques, 118
 - linguistic techniques, 105, 120
 - manually, 124
 - methods, 98
 - semantic networks, 109, 113, 116
 - using grouping techniques, 109
 - using techniques, 113
- category bar chart, 159–160
- category building, 6, 105, 109, 242
 - classification link exceptions, 113
 - co-occurrence rule technique, 6, 111, 123
 - concept inclusion technique, 123
 - concept root derivation technique, 6, 111, 123
 - semantic networks technique, 6, 111, 123
 - using techniques, 6, 111
- category name, 93
- category rules, 138–139, 144, 146–148
 - co-occurrence rules, 109, 113, 120
 - examples, 144
 - from concept co-occurrence, 6, 110–111, 114, 117, 123
 - from synonymous words, 6, 109–111, 113–114, 120, 123
 - syntax, 139
- category web graph/table, 159, 161–162
- changing
 - data source, 61
 - templates, 187
- charts, 159
- closed-ended question, 2
- cluster, 177

- co-occurrence rules technique, 6, 109, 111, 113–114, 117, 120, 123
- code frames, 126–127
- collision modifiers, 175
- colors
 - exclude dictionary, 223
 - for summary graph bars, 60
 - for types and terms, 210, 250
 - setting color options, 18
 - synonyms, 220
- column wrapping, 18
- combining categories, 152
- compact format, 132
- complete flag, 73
- componentization, 114
- concept inclusion technique, 6, 109, 111, 113–115, 120
- concept root derivation technique, 109, 113–114, 120, 123
- concepts
 - adding to categories, 100, 104, 150
 - adding to types, 87
 - best descriptors, 101
 - creating types, 84
 - excluding from extraction, 89
 - extracting, 77
 - forcing into extraction, 90
 - in categories, 100, 104
- coordinate systems
 - transforming, 173
- copying
 - categories, 156
- copying visualizations, 178
- Core library, 208
- creating
 - categories, 98, 105, 125
 - categories with rules, 139
 - category rules, 138–139, 146
 - exclude dictionary entries, 223
 - libraries, 196
 - optional elements, 220
 - projects, 27
 - synonyms, 84–85, 218
 - synonyms for Japanese, 250
 - template from resources, 186
 - type dictionaries, 209, 250
 - types, 87
- currencies (nonlinguistic entity), 228
- custom colors, 18

- data
 - categorizing, 91, 105, 124
 - category building, 6, 109, 111, 113, 120
 - changing data source, 61
 - data source selection, 28, 62
 - editing variable properties, 50
 - exporting, 52
 - extracting, 81
 - IBM SPSS Data Collection, 32, 54, 66
 - IBM SPSS Statistics *.sav* files, 29, 52, 54, 63
 - Microsoft Excel *.xls* / *.xlsx* files, 30, 52, 56, 64
 - ODBC, 32, 65
 - refining results, 84
 - refreshing, 71
 - sorting, 50
 - viewing, 49
- data pane, 95
 - display button, 94
- dates (nonlinguistic entity), 228
- deactivating nonlinguistic entities, 232
- default libraries, 195
- definitions, 100, 104
- deleting
 - categories, 157
 - category rules, 148
 - disabling libraries, 200
 - excluded entries, 223
 - libraries, 200, 202
 - optional elements, 221
 - resource templates, 188
 - synonyms, 221
 - type dictionaries, 216
- delimiter, 17
- dependency analysis, 239–240
- descriptors, 93
 - categories, 100, 104
 - choosing best, 101
 - editing in categories, 150
- dictionaries, 14, 207
 - excludes, 195, 207, 222
 - substitutions, 195, 207, 217
 - types, 195, 207
- digits (nonlinguistic entity), 228
- disabling
 - exclude dictionaries, 223
 - libraries, 200
 - nonlinguistic entities, 232
 - status bars, 74
 - substitution dictionaries, 221
 - synonym dictionaries, 227
 - type dictionaries, 216
- display button, 94
- display columns in the categories pane, 93
- display columns in the data pane, 95
- docs column, 93–94
- dodge, 177
- dollar sign (\$), 219
- drag and drop, 125

- e-mail (nonlinguistic entity), 228
- edit mode, 163
- editing
 - categories, 148–150
 - category rules, 147
 - properties, 149
 - refining extraction results, 84

-
- editing graphs
 - size of graphic elements, 167
 - editing visualizations, 163
 - adding 3-D effects, 173
 - automatic settings, 164
 - axes, 170
 - categories, 171
 - collapsing categories, 171
 - colors and patterns, 166
 - combining categories, 171
 - dashing, 166
 - excluding categories, 171
 - legend position, 178
 - margins, 168
 - number formats, 169
 - padding, 168
 - panels, 173
 - point aspect ratio, 167
 - point rotation, 167
 - point shape, 167
 - rules, 164
 - scales, 170
 - selection, 164
 - sorting categories, 171
 - text, 165
 - transforming coordinate systems, 173
 - transparency, 166
 - transpose, 173
 - empty responses, 26
 - enabling nonlinguistic entities, 232
 - entire project view, 10, 13
 - exclamation mark (!), 219
 - exclude dictionary, 195, 222–223
 - excluding
 - concepts from extraction, 89
 - disabling dictionaries, 216, 221
 - disabling exclude entries, 223
 - disabling libraries, 200
 - from category links, 113
 - from fuzzy exclude, 227
 - explore mode, 162
 - exporting, 52
 - as IBM SPSS Statistics *.sav* files, 54
 - as Microsoft Excel *.xls* / *.xlsx* files, 56
 - categorization results, 52–54, 56
 - for IBM SPSS Data Collection, 54
 - output formats, 53
 - predefined categories, 135
 - public libraries, 202
 - summary graphs, 58
 - templates, 189
 - extending categories, 120
 - extracting, 3, 77, 81–82, 195, 207, 237
 - concepts, types, and patterns, 77
 - forcing words, 90
 - nonlinguistic entities, 83
 - refining results, 8, 84
 - results, 25
 - saving results, 84
 - uniterms, 4, 83, 238
 - extraction patterns, 233
 - file recovery, 17
 - filtering libraries, 199
 - find and replace (advanced resources), 226
 - finding terms and types, 198
 - flagging responses, 73
 - flat list format, 131
 - font color, 210, 250
 - forced definitions, 233–234
 - forcing
 - concept extraction, 90
 - display columns, 95
 - force in, 95
 - force out, 95
 - responses, 153
 - terms, 214
 - words into categories, 154
 - frequency, 118
 - fuzzy grouping exceptions, 82, 225, 227
 - generate inflected forms, 207, 209–210, 250
 - global delimiter, 17
 - graphic elements
 - changing, 175
 - collision modifiers, 177
 - converting, 175
 - types, 175
 - graphs
 - category web graph, 159
 - editing, 163
 - explore mode, 162
 - exporting summary graphs, 58
 - refreshing, 159
 - size of graphic elements, 167
 - HTTP/URLs (nonlinguistic), 228
 - IBM SPSS Data Collection, 32, 66
 - changing data source, 61
 - exporting, 54
 - IBM SPSS Statistics *.sav* files, 29, 63
 - changing data source, 61
 - exporting, 54
 - output format, 53
 - ID variables, 25–26, 32, 66
 - ignoring concepts, 89
 - important flag, 73
 - importing
 - input data, 32, 66
 - ODBC, 31, 65

- predefined categories, 127
- preparing data, 26
- public libraries, 201
- refreshing data, 71
- reimporting data, 61
- templates, 189
- indented format, 133
- inflected forms, 114, 207, 209–210, 250
- IP addresses (nonlinguistic entity), 228

- Japanese, 237, 242
 - Resource Editor, 242
 - type properties, 250
 - types, 246, 252
- jitter, 178

- labels for categories, 104, 149
- language handling sections, 225, 233
 - abbreviations, 233–234
 - extraction patterns, 233
 - forced definitions, 233–234
- Language Weaver, 21, 34, 69
- legal notices, 255
- legend
 - position, 178
- **lib*, 201
- libraries, 14, 195, 207
 - adding, 197
 - Budget library, 208
 - Core library, 208
 - creating, 196
 - deleting, 200, 202
 - dictionaries, 195
 - disabling, 200
 - exporting, 202
 - importing, 201
 - library synchronization warning, 202
 - linking, 197
 - local libraries, 202
 - naming, 199
 - Opinions library, 208
 - public libraries, 202
 - publishing, 51, 204
 - renaming, 199
 - sharing, 73
 - sharing and publishing, 202
 - shipped default libraries, 195
 - synchronizing, 202
 - updating, 204
 - viewing, 199
- linguistic resources, 25, 195
 - resource templates, 183
 - text analysis packages, 40, 42
- linguistic techniques, 3, 6, 8, 111
- link exceptions, 113
- Location type dictionary, 208

- making templates from resources, 186
- managing
 - categories, 148
 - local libraries, 199
 - public libraries, 201
- marking responses, 73
- match option, 207, 209–210, 212, 250
- maximum number of categories to create, 112
- maximum search distance, 112, 117, 123
- merging categories, 152
- Microsoft Excel *.xls* / *.xlsx* files, 30, 56, 64
 - changing data source, 61
 - exporting predefined categories, 135
 - importing predefined categories, 126–127
 - output format, 53
- minimum link value, 112
- monitoring, 73
- moving
 - categories, 151
 - type dictionaries, 216
- muting sounds, 20

- naming
 - categories, 104, 149
 - libraries, 199
 - type dictionaries, 215
- natural sort, 50
- Negative type dictionary, 208
- new categories, 124
- new features, 1
- nonlinguistic entities, 83
 - addresses, 228
 - amino acids, 228
 - currencies, 228
 - dates, 228
 - digits, 228
 - e-mail addresses, 228
 - enabling and disabling, 232
 - HTTP addresses/URLs, 228
 - IP addresses, 228
 - normalization, *NonLingNorm.ini*, 231
 - percentages, 228
 - phone numbers, 228
 - proteins, 228
 - regular expressions, *RegExp.ini*, 229
 - times, 228
 - U.S. social security number, 228
 - weights and measures, 228
- normalization, 231
- NOT rule operator, 147

- ODBC data source, 31–32, 61, 65
- open-ended question, 2
- opening projects, 47
- operators in rules & | !() , 147
- Opinions library, 208

- optional elements, 217
 - adding, 220
 - definition of, 218
 - deleting entries, 221
 - target, 220
- options, 16
 - display, 18
 - sound, 20
 - system, 17
 - translation, 21
- OR rule operator, 147
- Organization type dictionary, 208

- palettes
 - displaying, 164
 - hiding, 164
 - moving, 164
- part-of-speech, 233–234
- patterns, 77, 80
- percentages (nonlinguistic entity), 228
- permutations, 83
- Person type dictionary, 208
- phone numbers (nonlinguistic), 228
- plural word forms, 210
- polar coordinates, 173
- Positive type dictionary, 208
- predefined categories, 126–127, 135
 - compact format, 132
 - flat list format, 131
 - indented format, 133
- preferences, 16, 18, 20
- preparing your data, 26
- printing categories, 156
- Product type dictionary, 208
- projects, 25, 47
 - creating, 27
 - data source, 28, 62
 - opening, 47
 - options for libraries, 17
 - properties, 48
 - renaming, 51
 - reusing categories, 156
 - saving, 51
 - selecting categories and resources, 36
 - sharing, 73
 - status bar, 74
 - text analysis packages, 36
 - translation, 34, 69
 - variable selection, 32, 66
- properties
 - categories, 104, 149
 - for Japanese types, 250
 - projects, 48
 - variables, 50
- proteins (nonlinguistic entity), 228
- publishing, 51, 204
 - adding public libraries, 197
 - libraries, 202
- punctuation errors, 82

- question view, 10–11

- records, 95
- recovered files, 17
- reference variables, 25–26, 32, 66
- refining results
 - adding concepts to types, 87
 - adding synonyms, 85
 - categories, 8, 148
 - creating types, 87
 - excluding concepts, 89
 - extraction results, 8, 84
 - forcing concept extraction, 90
- refreshing graphs, 159
- reimporting data, 61
- relevance of responses and categories, 96–97
- reliability, 8
- renaming
 - categories, 124, 149
 - libraries, 199
 - projects, 51
 - resource templates, 188
 - type dictionaries, 215
- replacing resources with template, 187
- reports and summary graphs, 58
- resource editor, 10, 186–187, 225
 - for Japanese, 242
 - global delimiter option, 17
 - making templates, 186
 - switching resources, 187
 - updating templates, 186
- resource templates, 4, 183, 238
- resources
 - backing up, 190
 - editing advanced resources, 225
 - restoring, 190
 - shipped default libraries, 195
 - switching template resources, 187
- responses, 95
 - flagging, 73
 - forcing into categories, 153
 - marking as complete, 73
- restoring resources, 190
- reusing
 - categories, 156
- rules
 - Boolean operators, 147
 - co-occurrence rules technique, 117
 - creating, 146
 - deleting, 148
 - editing, 147
 - syntax, 139

- saving
 - autosave of projects, 17
 - extraction results, 84
 - projects, 51
 - resources, 190
 - resources as templates, 186
- score button, 94
- scoring, 94
- secondary analysis
 - dependency analysis, 239
 - sentiment analysis, 239
- semantic networks technique, 6, 109, 111, 113–114, 116, 120, 123
- sentiment analysis, 239–240
 - options, 240
- separators, 17
- settings, 16, 18, 20
- sharing libraries, 202
 - adding public libraries, 197
 - publishing, 51, 204
 - updating, 204
- sharing projects, 73
- shipped (default) libraries, 195
- social security # (nonlinguistic), 228
- sorting data and variables, 50
- sound options, 20
- spelling mistakes, 26, 82, 227
- stack, 177
- statistics
 - descriptions, 175
 - editing in visualizations, 175
- status bars, 74
- storing extraction results, 84
- substitution dictionary, 195, 217, 220–221
- summary graphs, 58
- survey data, 2, 25–26, 32, 66
- synchronizing libraries, 202, 204
- synonyms, 84, 217
 - ! ^ * \$ symbols, 219
 - adding, 85, 218, 250
 - colors, 220
 - definition of, 217
 - deleting entries, 221
 - for Japanese text, 250
 - fuzzy grouping exceptions, 82, 227
 - target terms, 218, 250
- *.tap text analysis packages, 36, 40–42, 44
- target terms, 220
- techniques
 - co-occurrence rules, 109, 113, 117, 120
 - concept inclusion, 109, 113, 115, 120
 - concept root derivation, 109, 113–114, 120
 - drag and drop, 125
 - frequency, 118
 - semantic networks, 109, 113, 116, 120
 - templates, 4, 183, 238
 - backing up, 190
 - deleting, 188
 - importing and exporting, 189
 - making from resources, 186
 - renaming, 188
 - restoring, 190
 - switching templates, 187
 - TLA, 187
 - updating or saving as, 186
 - term componentization, 114
 - terms
 - adding to exclude dictionary, 223
 - adding to Japanese types, 244
 - adding to types, 210
 - color, 210, 250
 - finding in the editor, 198
 - forcing terms, 214
 - forcing words into categories, 154
 - inflected forms, 207
 - match options, 207
 - text analysis, 3, 7–8, 10
 - text analysis packages, 36, 40–42
 - text match, 104, 149, 154
 - text mining, 3
 - text separators, 17
 - text variables, 25–26, 32, 66
 - times (nonlinguistic entity), 228
 - titles for the export summary graph, 60
 - TLA, 187
 - tracking responses, 73
 - trademarks, 256
 - translation, 21, 71
 - options dialog, translation tab, 21
 - translating into English, 34, 69, 71
 - translation accuracy setting, 35, 70, 72
 - translation settings dialog, 71
 - type dictionary, 195
 - adding terms, 210
 - adding terms for Japanese, 244
 - built-in types, 208
 - creating types, 209, 250
 - deleting, 216
 - disabling, 216
 - forcing terms, 214
 - moving, 216
 - optional elements, 207
 - renaming, 215
 - synonyms, 207
 - type frequency, 118
 - types, 207
 - adding concepts, 84
 - built-in types, 208
 - creating, 209, 250
 - default color, 18, 210, 250
 - dictionaries, 195
 - extracting, 77

-
- finding in the editor, 198
 - for Japanese, 246, 250, 252
 - type frequency, 118
- uncategorized, 93
- Uncertain type dictionary, 208
- uniterms, 83
- Unknown type dictionary, 208
- updating
 - graphs, 159
 - libraries, 202, 204
 - templates, 186
- variables
 - changing data source, 61
 - editing properties, 50
 - exporting, 52–53
 - ID variables, 25–26, 32, 66
 - importing, 29, 31, 63, 65
 - matching, 68
 - reference variables, 25–26, 32, 66
 - refreshing, 71
 - text variables, 25–26, 32, 66
- viewing
 - categories, 159
 - data, 49
 - libraries, 199
- views
 - entire project, 13
 - question view, 11
 - resource editor window, 14
 - text analysis window, 10
- visualization pane
 - category web graph, 159
 - updating graphs, 159
- visualizations
 - axes, 170
 - categories, 171
 - colors and patterns, 166
 - copying, 178
 - dashings, 166
 - editing, 163
 - legend position, 178
 - margins, 168
 - number formats, 169
 - padding, 168
 - panels, 171, 173
 - point aspect ratio, 167
 - point rotation, 167
 - point shape, 167
 - scales, 170
 - text, 165
 - transforming coordinate systems, 173
 - transparency, 166
 - transpose, 171, 173
- web graphs, 159
- web table, 159
- weights/measures (nonlinguistic), 228
- what's new, 1
- .xls / .xlsx files, 53, 56*