



Topic Factory Developer's Guide

IBM ViaVoice™ SDK for Windows®

Version 8.0

Printed in the USA

Note

Before using this information and the product it supports, be sure to read the general information under Appendix C, "Notices."

Second Edition (May 2001)

The following paragraph does not apply to the United Kingdom or any country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF 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 publication 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.

It is possible that this publication may contain reference to, or information about, IBM products (machines and programs), programming, or services that are not announced in your country. Such references or information must not be construed to mean that IBM intends to announce such IBM products, programming, or services in your country.

Requests for technical information about IBM products should be made to your IBM reseller or IBM marketing representative.

©Copyright International Business Machines Corporation 1999, 2001. All Rights Reserved.

Note to U.S. Government Users—Documentation related to restricted rights— Use, duplication or disclosure is subject to restrictions set forth in GS ADP Schedule Contract with IBM Corp.

Contents

About this Book		5
	Who Should Read This Book	5
	Related Publications	5
	Conventions in this Book	6
Chapter 1	Introducing Topic Factory	7
	What is a Topic?	7
	What is Topic Factory?	8
	Developer Skills	9
	Installing Topic Factory	11
	System Requirements	11
Chapter 2	Getting Started	13
	Preparing to Build a Topic	13
	Gathering Data	13
	Cleaning Data	14
Chapter 3	Using Topic Factory	15
	Developing a Topic	16
	Naming and Describing Your Topic	16
	Selecting Files	18
	Choosing Topic Words	20
	Adding Words to Your Topic	22
	Create and Edit Baseforms	24
	Flags File (optional)	25
	Completing Your Topic	26

Chapter 4	Testing Your Topic	27
	Installing Your Topic for Testing	27
	Activating Your Topic	27
	Testing for Quality	28
	Testing for Recognition Accuracy	28
	Reviewing and Evaluating your Test Results	29
	Viewing your Tokenized Files	29
	Additional Pronunciations	31
	Improving Topic Accuracy	33
	Completing Your Topic	33
Chapter 5	Packaging Your Topic for Distribution	35
	Creating Your Help File	35
	Integrating Your Help File	36
	Distributing Your Topic	36
Appendix A	German-Language Topic Building	37
	Gathering Data	37
	Tokenizing	37
	Lowercasing of Uppercase-Only Words	37
	Abbreviations	38
	Installation	38
Appendix B	U.S English Baseforms	39
Appendix C	Notices	43
	Trademarks	44
Index		45

About This Document

The IBM Topic Factory Developer's Guide provides the information you'll need to build topics, specialized vocabularies, for IBM ViaVoice™ products. The Topic Factory tool is designed to be used by ISVs (Independent Software Vendors) and VARs (Value Added Resellers). This book is prepared in Portable Document Format (PDF) to provide the advantages of text search and cross-reference hyperlinking and is viewable with the Adobe Acrobat Reader v.3.x. We recommend that you print all or part of this guide for quick reference.

Who Should Read This Book

This guide is written primarily for vocabulary and topic developers. Developers should have the skills required to gather source data, a basic knowledge of programming to process source data, a familiarity with Windows® 95/98 or Windows NT™, a basic understanding of phonetics, an aptitude for understanding language models, and some expertise in the area of the topic to be built. Read more about the knowledge required to build a topic in [“Developer Skills” on page 9](#).

Related Publications

Refer to the following sources for additional information, frequently asked question and technical support:

- IBM ViaVoice Developer's Corner website at:
<http://www.ibm.com/software/speech/dev>
Click on Topic Factory

Conventions in this Book

The following conventions are used to present information in this book:

<i>Italic</i>	Used for emphasis and for references to other documents.
Bold	Represents a menu option or other user interface control, such as command buttons.
<code>Courier Regular</code>	Represents text from your source data or source files.

The IBM Topic Factory is the tool you need to build topics for ViaVoice products. Keep reading, and you'll find out more about topics and how to create topics with Topic Factory.

What is a Topic?

Topics are small, specialized vocabularies designed to enhance recognition accuracy for ViaVoice products. When used in conjunction with ViaVoice dictation vocabularies, Continuous General Dictation, or an industry vocabulary, such as the IBM ViaVoice Medical Vocabulary—topics add relevant terminology and enhances recognition accuracy. For example, a doctor whose practice includes knee and sports medicine would activate the IBM ViaVoice Medical Vocabulary and an orthopedic topic. ViaVoice users can select as many relevant topics as they want from ViaVoice Options. A topic contains the following:

- A set of words
- Pronunciations for the words
- A language model

A language model consists of statistics that are used in predicting word sequences, which words are likely to follow one another. A topic language model is considerably smaller than a full-size vocabulary language model. The size of a topic makes it easy to download from a Web site.

Topics are separate from a user's Personal Language Model or cache. Topics are read-only and are accessible by all users of ViaVoice. Topics may contain up to 65,536 words (64K). This is in addition to the 64,000 words in the full-size vocabulary.

What is Topic Factory?

The Topic Factory user interface is designed as a wizard and is simple to use. It easily guides you through the following steps to build a ViaVoice topic:

- Naming and describing a topic
- Selecting the source files and parsing the data
- Choosing the words to add to your topic
- Training pronunciations for new words
- Editing baseforms (pronunciations)
- Building an installable topic file

Topic Factory is a scaled-down version of the tools that are used to produce full-size vocabularies, such as the IBM ViaVoice Medical Vocabulary and the IBM ViaVoice Legal Vocabulary. Topic Factory is not designed to be used by end-users. It is designed to be used by developers who have the knowledge, skills, and aptitude to create a topic. The following section describes in detail the skill level required for building topics.

Developer Skills

Specialized skills are required to build a topic. The source data should be prepared before using Topic Factory, and a level of knowledge and expertise is needed in the subject area of the topic being built. The following skills are required to create a topic:

Understanding of Windows 95/98 or Windows NT

The topic developer must have the basic understanding of Windows 95/98 or Windows NT, including removing files, checking available disk space, and viewing the contents of files.

Understanding of ViaVoice

The developer must have an understanding and a working knowledge of ViaVoice. These skills will be used when testing and reworking your topic.

Programming Skills

The developer must know how to gather the appropriate source data and have the skills to process that data. For example, control characters, headers, and footers need to be removed from the source data before adding the files to Topic Factory. The developer must be able to write simple programs to remove this extraneous information.

Understanding of Language Models

The developer must have an aptitude for learning about bigrams and trigrams and how context influences the recognition process.

Expertise in the Domain of the Topic Subject

The developer, or someone working with the topic developer, must be able to identify words that are frequently used in the domain of the topic being built. An understanding of special word formatting is also required. For example, words in the legal domain can be Latin phrases, such as “de facto,” or abbreviations with special pronunciations, such as “A.2d”, pronounced A-second.

Speaking Skills

To produce accurate pronunciations when training new words, the developer should be a native speaker of U.S. English, without any strong regional accent. Editing pronunciations generated by Topic Factory requires knowledge of the subtleties of how phonemes differ from one another. Some of these differences are not present in other languages, especially the vowel sounds.

Understanding of Phonetics

The developer must have an aptitude for learning the phonetic representation of pronunciations, such as those given in a dictionary. When editing baseforms, the developer must be able to write the word with its pronunciation symbols. For example, the baseform for “snails” is “S N EY L Z .”

Note:

The developer is not required to have all these skills. For example, a developer on a team may have the programming skills required to build the topic, but may need the expertise of someone who understands the content of the topic.

Installing Topic Factory

Your Topic Factory software package includes a CD-ROM containing the installer software, the program files, and this online guide.

Before installing Topic Factory make sure you have sufficient memory and disk space, and that no programs are running that would interfere with installation.

System Requirements

The Topic Factory requires the following system configuration:

- Windows® 95/98 or Windows NT™ 4.0
- Processor performance equivalent to Intel® Pentium® 166MHz with MMX and 256K L2 cache
- 64 MB of RAM
- Sufficient free hard disk space
 - 2 MB for Topic Factory programs
 - You will need 3 times the total file size of your source data. For example, if you are processing 100 MB of source data, you need 300 MB of free hard disk space
- CD-ROM drive
- IBM ViaVoice™ Millennium or higher installed

Install Topic Factory

1. Insert the Topic Factory CD-ROM in your CD drive to start the installation program.
2. Follow the instructions on the screen.

Note:

If the installation program does not automatically start, click **Start** on the Windows taskbar, and then click **Run**. Type d:\setup (where d: is the letter of your CD-ROM drive).

Uninstall Topic Factory

The Uninstall program for Topic Factory is located in the **IBM ViaVoice Topic Factory** folder. To uninstall Topic Factory, click the **Start** button > **Programs** > **IBM ViaVoice Topic Factory** folder and then click **Uninstall ViaVoice Topic Factory**.

The quantity and quality of the data you collect affects the overall quality of the topic you build. Follow the suggestions for Gathering Data and Cleaning Data, and you will be on your way to creating a high-quality topic.

Preparing to Build a Topic

Before you begin preparing and gathering data, consider the topic you're going to build. You have to specifically understand the language domain of your topic audience, and you must limit the scope of your topic. For example, you would not want to do a medical topic; that would be too broad a scope. Specialized medical practices are ideal for topics, such as Cardiology, Chiropractic, Oncology, Ophthalmology, Orthopedics, and Pediatrics. For the best speech recognition, the user can activate any number of topics with a full-size medical vocabulary, although at most, two is recommended for better performance. For example, in the medical domain, a Sports Injury specialist might want topics on "Fitness" and "Nutrition" to use with a medical base vocabulary.

Gathering Data

Your first step in building a topic is gathering the source files to be analyzed by Topic Factory. This is a *very* important step because your topic's ability to recognize words begins with the quality of your source data. Consider the following when gathering data:

Text

The text in your data should be representative of real writing, matching as closely as possible the dictation of your topic audience. This ensures words are added with the context in which they are used. ViaVoice uses context information to improve speech recognition accuracy.

Source

Use as many sources and as much data as possible. For example, a topic built on data from one doctor's office would not be representative of wide-scale medical practices.

The source data should be focused on the scope of your specific topic. For example, you would not use general litigation text to create a topic for Tax Attorneys—the data would not be specific enough, nor would you use correspondence from only one company to create an e-mail topic—the data would be too specific.

Format

The source files must be in ASCII text format (.txt). The source data should not be in all uppercase letters (if the data is in all uppercase letters, almost every word will be displayed as **Not in Vocabulary**). See [“Choosing Topic Words” on page 22](#) for information on words appearing in the **Not in Vocabulary** list box.

Size

It is recommended to have a minimum of two million words in the source data. The amount of data depends on the domain of the topic. The broader the domain, the more data that will be needed.

Cleaning Data

The “cleaner” your source data is, the easier it is to add words to your topic. Remove headers, footers, control characters, tables, and other extraneous characters before analyzing your text files.

If you do not remove these characters, they show up as words you can choose to add to your topic. For example, if your source data contains headers with initial caps, each word with an initial capital will be identified as **Not in Vocabulary** because these words are not in the base vocabulary or back-up dictionary with capital letters. For this reason, you should remove the headers. Not cleaning your data could result in having to sort through many words you do not want to include in your topic.

Data that contains extraneous text, such as slang and typos, is considered “dirty” when running Topic Factory. For example, the data typically found in newsgroup forums is dirty because it contains grammatical errors, slang words, and an overwhelming number of misspellings. If the data is too dirty, do not include it in your source files.

Note:

Keep a small file of some of the data you have prepared to use later for testing your new topic. Make sure this data is not included in the data you use to build your topic. For more information on using this data, see [“Testing for Recognition Accuracy” on page 30](#).

Before starting Topic Factory, add a new user on the **User** page in **ViaVoice Options**. You must add a new user so there are no personal vocabulary files that may interfere with the building of your topic. Also on the User page, select the base vocabulary you want activated when building your topic (Continuous General Dictation or an industry vocabulary). Do not activate any topics. To start **Topic Factory**, click the **Start** button > **Programs** > **IBM ViaVoice Topic Factory** folder and then click **Topic Factory**..

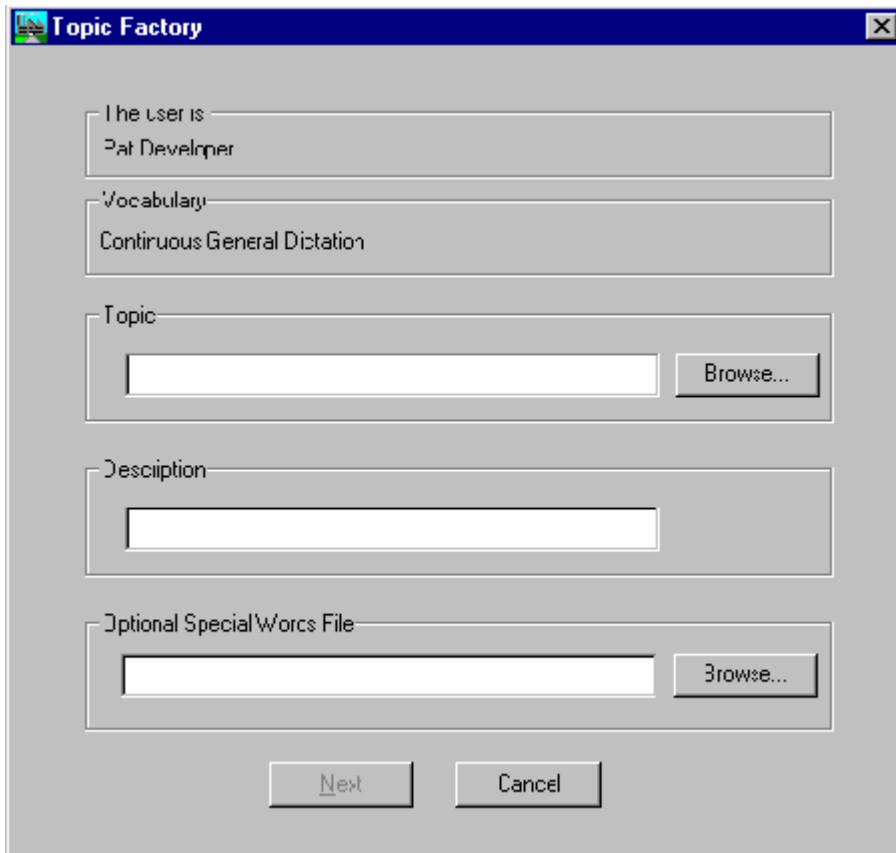


Figure 1. Topic Factory Dialog

Developing a Topic

The following steps take you from naming your topic through building your installable topic.

Naming and Describing Your Topic

The following describes the three text boxes on the Topic Factory dialog:

Topic

In the **Topic** box, type the name of your topic. A topic name is used to name all your topic files. The name should be one word with a maximum of six characters. For example, the name for a medical topic called Family Practice could be FPMED.

Description

In the **Description** box, type the words that describe your topic. This description appears on the **User** page of **ViaVoice Options**, and is the name the user chooses when activating your topic in the **Topics** list box. For example, a description for a financial business topic might be “Business and Finance.”

Optional Special Words File

In the **Optional Special Words File** box, type a name for an optional file (.txt extension). This file is used for the list of words or phrases that get parsed incorrectly. The following is a list of reasons why you would want to create this file:

- **Tokenization**

Topic Factory has predefined rules for parsing (separating) words and punctuation by delimiting with spaces. For example, the sentence

John said, “On December 5, 1998, the cross-country team will be in Santa Fe, New Mexico.”
would appear like this after tokenizing:

John said , “ On December 5 , 1998 , the cross - country team will be in Santa Fe , New Mexico . ”

The tokenization rules are not always going to allow your data to be parsed the way you want. For example, you may find the tokenizer splits “TCP/IP” into “TCP / IP”. If you think a word is going to be parsed incorrectly, add it to your **Optional Special Words File**.

- **Abbreviations**

Topic Factory might not know that a period in an abbreviation is not a period at the end of a sentence. For example, it may tokenize an abbreviation such as “tsp.” into “t sp . ”. To make sure Topic Factory treats your abbreviations as one word, you can add these words to your **Optional Special Words File**.

- **Phrases**

You can add phrases to the **Optional Special Words File** such as “New York”. If the tokenizer parses your phrase into separate words, it will be confusing when they appear in the Not in Vocabulary list box. For example, the phrase “pasta e fagiole” would appear as three separate words. You would have to record each part of this phrase as a separate word. When you put a phrase in your **Optional Special Words File**, the tokenizer does not split it and allows you to record the words in the phrase together.

- **Casing**

You can expect problems in casing when tokenizing. For example, if your data contains “CAT scan”, you might find that CAT is lowercased to “cat”. To correct this, add CAT to your **Optional Special Words File**.

- **Hyphenated Words**

Hyphenated words might be split during tokenization. For example, “X-Ray” could be tokenized into “X - Ray”; this requires the topic user to say “X DASH Ray” when dictating. If you add frequently used hyphenated words to your **Optional Special Words file**, the user can dictate hyphenated words without saying “DASH” between words.

- **Dictionary File**

The **Optional Special Words File** could also be a domain-specific list of words, serving the function of a dictionary when tokenizing your data. For example, an extensive list of medical terms might be used when building a medical related topic.

Note:

Remember, you can only specify one file in the **Optional Special Words File** field, but you can update the file as often as necessary while building and testing your topic.

Selecting Files

Once you complete the **Topic Factory** dialog box, the **Topic Factory: Select Files** dialog box appears.

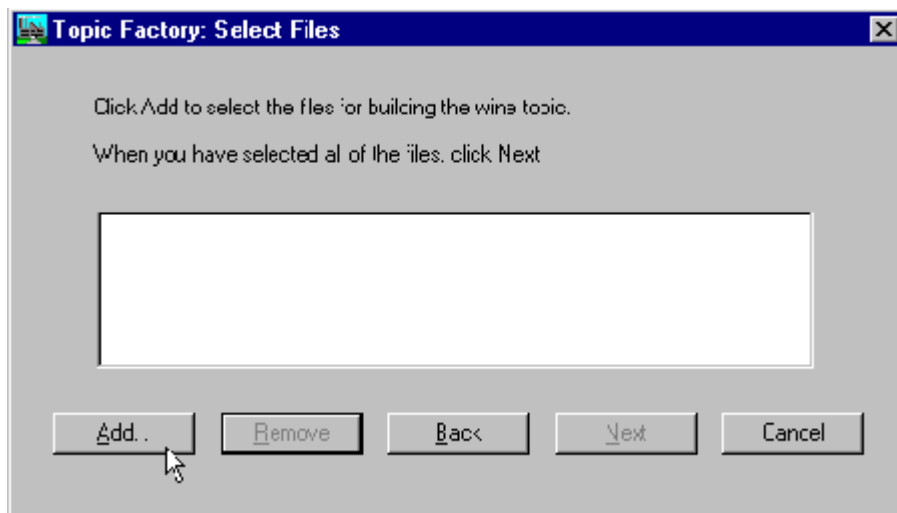


Figure 2. Select Files Dialog

When selecting files, remember to add only clean files. For more information on clean files, see [“Cleaning Data” on page 14](#). To select files for building your topic, Click **Add** and use the Windows **Open** dialog box. To select contiguous files, select the first file, and press **Shift** when selecting the last file. To select non-contiguous files, press **Ctrl** when selecting each file.

Use the **Add** button repeatedly to add files from multiple directories. There is no limit to the size or number of the files you add.

Topic Factory remembers the files you add to your topic; if you exit Topic Factory, the files you selected in a previous session are used when you restore the session.

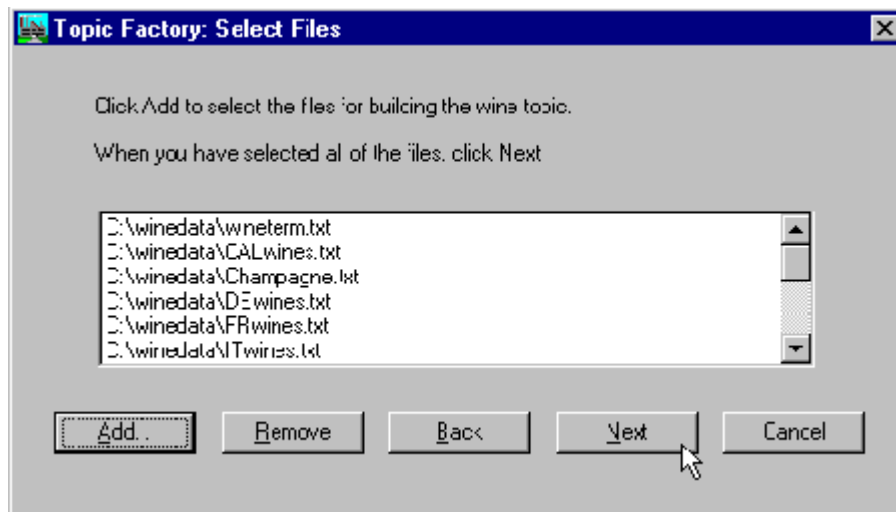


Figure 3. Select Files Dialog with added files

When you click **Next**, a progress indicator will appear which will be the process of analyzing your data. The time it takes for Topic Factory to tokenize your data depends on the size of your source files, the speed of your processor, and the amount of memory.

You must reparse your data every time you update your **Optional Special Words File**, select new files to add to your topic, or change data in a selected file. To reparse your data, answer **Yes** when prompted. You are prompted to reparse when you open an existing topic and click **Next** on the **Select Files** dialog box.

The next step is choosing words to add to your topic.

Choosing Topic Words

When Topic Factory completes tokenizing your data, the **Topic Factory: (Name of your topic)** dialog box appears.

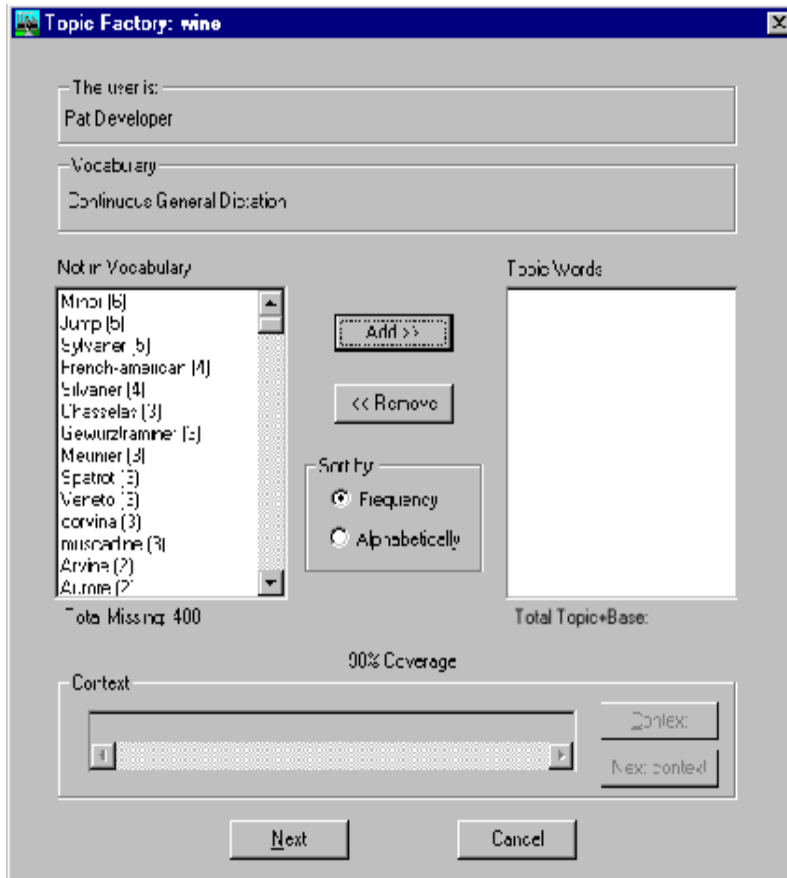


Figure 4. Topic Factory Dialog

All the words in your source data that were in the current full-size vocabulary automatically become part of your topic. Words from your source data that were not found in your currently active full-size vocabulary appear in the **Not in Vocabulary** list box. The number in parentheses after each word is the

number of times the word occurred in your source data. You can sort this list by this frequency number or sort alphabetically.

There are several things to consider when choosing words for your topic:

- **Frequency Counts**

Use the frequency counts as a guideline in deciding which words to add to your topic. It's important to add the most frequently occurring words to your topic, but do not ignore lower frequency count words, as they may be just as important.

Frequency counts are a good indicator of the quality of your source data. If you find a "good" word to add to your topic that has a low frequency count, it might be an indicator to add additional source data to increase the amount of context for this word.

Do not automatically add a group of words with the highest frequency counts. Check each word individually, or you might add words you do not want in your topic.

- **Misspelled Words**

This is where it's important to have some expertise in the subject area of your topic. Do not automatically add, for example, the top 50% of the words with the highest frequency counts.

Many words in your data can be spelled incorrectly, and the word you want can be misspelled several times. For example, a word like "scaloppine" could be spelled six different ways, and the top three spellings could be incorrect. You must carefully look at each word before adding it to your topic.

- **Hyphenated Words**

You might find hyphenated words in your source data have been split into two words in the **Not in Vocabulary** list box. If this happens, you must go back and add the hyphenated word to your **Optional Special Words File**.

- **Casing**

If you see the same word written in multiple casings, make sure you select only the correct spelling. For example, do not add the word "The", but you might see the word "Brown". If you view the word in context (see "[Total Topic+Base](#)" on page 25), you'll find it belongs to "Dr. Brown" and you'll want to add it to your topic.

Adding Words to Your Topic

Select the words in the **Not in Vocabulary** list box and click **Add** to add words to your topic.

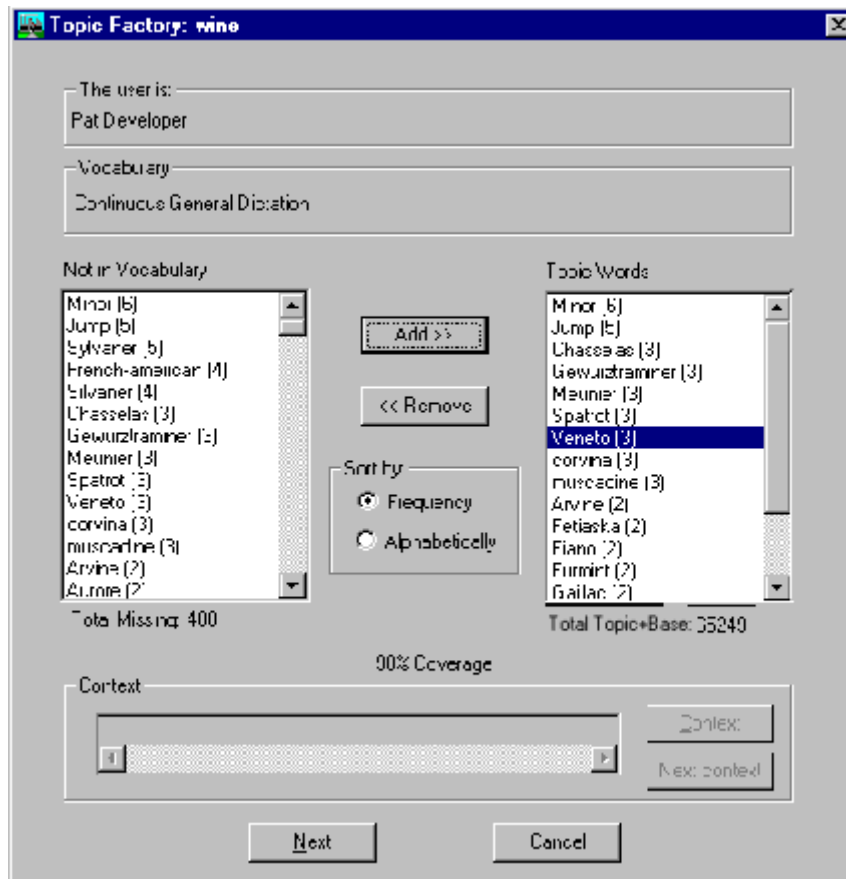


Figure 5. Topic Factory Dialog with Topic Words added

Total Topic+Base

The **Total Topic+Base** number indicates the total number of words in the topic. This number is the sum of the base vocabulary words found in your text plus the number of missing words you added to your topic. This number will increase as you add words to the **Topic Words** list box. You must ensure that this number does not exceed 64K.

Coverage

As you continue to add words to the **Topic Words** list box, the **% Coverage** indicator shows the percentage of coverage based on your source data. This percentage includes words that are already in your full-size vocabulary, plus the words selected for your topic. For example, the percentage indicator might show 70% before you add any words to your topic. This means that 70% of the words in your source data are already in your topic. As you add words to your topic, the percentage indicator increases.

Viewing a Word in Context

Viewing a word in context helps you understand why the word appears as it does, helps you decide if it is a “good” word or if it was tokenized correctly. If you need assistance in choosing a word, highlight the word and click the **Context** button. Sentences from your source data, in which the word appears, are displayed in the **Context** field. You can view up to 30 examples in context by clicking the **Next context** button. The context of a word is displayed by selecting a word in either the **Not in Vocabulary** list box or the **Topic Words** list box.

Note:

If you click **Cancel**, Topic Factory restores your words in the **Topic Words** list box when you resume.

Create and Edit Baseforms

Topic Factory analyzes the words you chose to add to your topic. It checks every word to see if it is in the 260,000 word backup dictionary. Words that appear in the backup dictionary have a baseform (a phonetic representation of a pronunciation) and do not require training. Words that do not appear in the backup dictionary, require training—recording a pronunciation—to produce baseforms. You must have an understanding of how words are phonetically represented to edit baseforms. It is important to review the baseforms of the words you are adding to your topic to be sure the pronunciations are correct.

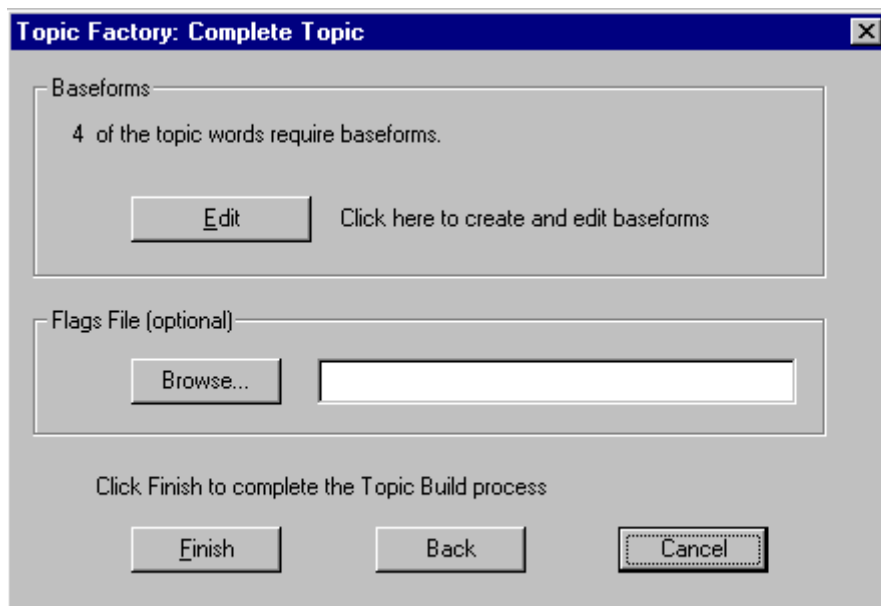


Figure 6. Topic Factory: Complete Topic Dialog

To create and edit the words that require baseforms, click the **Edit** button in the Baseforms list box. The Dictionary Builder will automatically load and search for the pronunciations of the topic words.

Note:

For more information, refer to the IBM Dictionary Builder help.

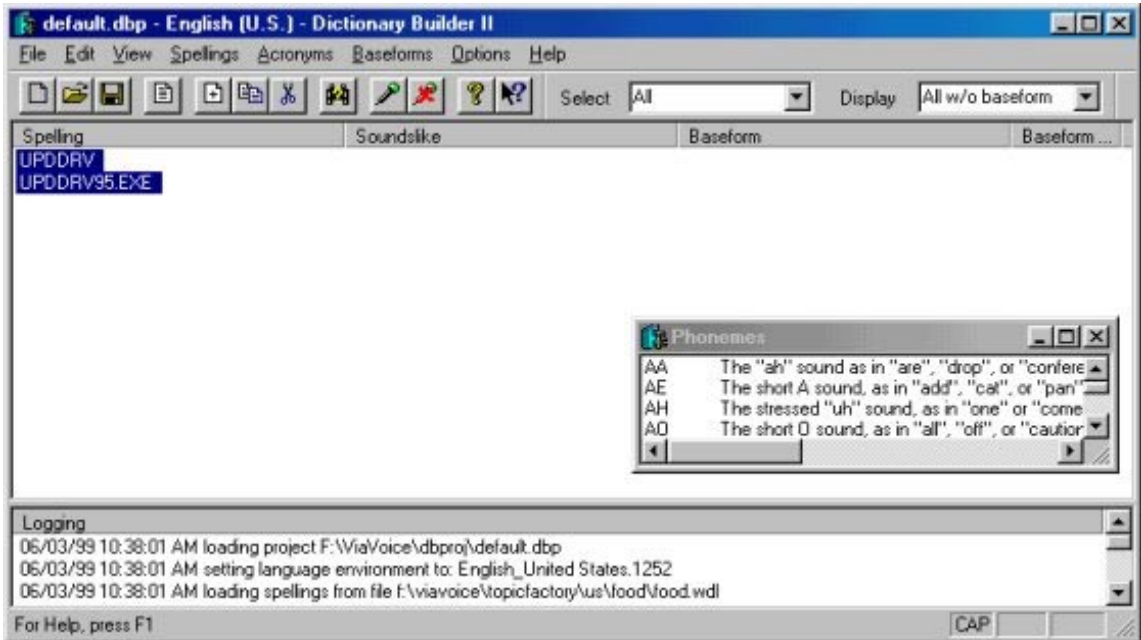


Figure 7. Dictionary Builder Dialog

You must create a pronunciation for those words that don't have one. If you want to add a pronunciation for a word in the base vocabulary, this can be done by doing a display all and editing.

Note:

If you have a database of pronunciations that you would like Dictionary Builder to use, put it in the “\viavoice\...\pools” directory and it will be automatically loaded. For more information on Loading a Pronunciation Pool, refer to the IBM Dictionary Builder help.

Flags File (optional)

If some of the topic words require special formatting, you may specify this via the **Flags File (optional)**. The format of the file is a word or phrase and its flags per line. For example, your **Flags File** could look like this:

- .com -L
- tsp. -A
- Professor -C

- € - D

The following flags can be specified in a topic:

- C = capitalize (the word that follows will be capitalized)
- L = glue left (glue this word to the preceding word)
- R = glue right (glue this word to the following word)
- D = digit (glues this word to any digits preceding or following the word)
- A = abbreviation (this word is an abbreviation. Makes it possible to use Spell-out with this word.)
- U = user defined (A speech enabled program can apply its own formatting when a word with this flag is decoded.)

Completing Your Topic

When you have finished editing all the baseforms, you are ready to complete the Topic Factory build process—building of the recognition engine data files.

In the **Topic Factory: Complete Topic** dialog, click **Finish**. You are prompted to include a filename for your help file. Type your topic name with an .hlp or .txt file extension.

Initially, the help file (.hlp or .txt) can be a “blank” file. After you are done building your topic, follow the guidelines for developing your help file in [“Packaging Your Topic for Distribution” on page 37](#).

To ensure the quality of your topic, you must thoroughly test your topic. Detailed testing information is in [“Testing Your Topic” on page 29](#).

Testing and reworking your topic are very important to the overall quality of your finished topic. Before you begin testing and reworking, your topic must be installed and activated.

Installing Your Topic for Testing

To test your topic on the computer, on which your topic was built, use ViaVoice **Vocabulary and Topic Installer** and install your topic from the directory where your topic files are located:

- `viavoice\topicfactory\us\[topic name]\release`

Open the **Vocabulary and Topic Installer** from the **IBM ViaVoice** folder and follow the on-screen instructions.

Activating Your Topic

To activate your topic, open **ViaVoice Options**. Click the **Start** button > **Programs** > **IBM ViaVoice**. From the **VoiceCenter**, click the **ViaVoice** menu > **User Options** > **ViaVoice Options** > select the **User** tab and in the **Topics** list box, select the name of your topic.

Testing for Quality

Testing is an important step to ensure the quality of your topic. It is an iterative process that includes evaluating your data each time you build your topic. The steps of the testing process include:

- Testing for recognition accuracy
- Reviewing and Evaluating your test results
- Viewing your tokenized data
- Updating your **Optional Special Words File**
- Reparsing your data
- Editing baseforms

The following suggestions for testing and reviewing your topic are recommended.

Testing for Recognition Accuracy

Use the data file you saved for testing (see [“Cleaning Data” on page 14](#)) as a script for speech recognition accuracy testing. The following testing scenario provides you with a measurement of accuracy for your topic:

1. Create a new user in **ViaVoice Options**. You must add a new user so there is no personal vocabulary or personal language model which might contain words or context information that would affect your test results.
2. Make sure your topic is activated for the new user.
3. Use **SpeakPad** to dictate your test script. To open **SpeakPad** from **ViaVoice VoiceCenter**, make sure your microphone is on, and say Dictate to SpeakPad.
4. Dictate your test script, remembering to say punctuation marks. Dictate 50 to 100 sentences for a reliable measurement.
5. Compare the recognition of your dictated speech to the original test script. You calculate your accuracy percentage by dividing the number of correctly recognized words by the total number of words dictated. For example, if you dictate 500 words and get 30 recognition errors, (500 minus 30 divided by 500) your speech recognition accuracy is 94%.

Reviewing and Evaluating your Test Results

Review your dictated accuracy test file to determine the cause of speech recognition errors. Here are a few suggestions when reviewing your file:

- Listen to the audio associated with each recognition error by selecting the misrecognized word and clicking **dictation > playback**. This ensures that you pronounced the word correctly. If you want to verify that this word was added to your topic, check the **Topic Words** list box.
- Notice if a word you added to your topic is consistently misrecognized. If so, go back and check the baseform pronunciation; you might need to edit the baseform.
- If a phrase is consistently being misrecognized, you might want to add it to your **Optional Special Words File**. Remember, you must reparse your data every time you update the **Optional Special Words File**.

When evaluating test data, do not correct misrecognized words using the **Correction** window. This would update the Personal Vocabulary of the test user, and would affect future test results.

For comparative analysis, repeat the accuracy testing with several speakers. Each tester should use a different user name and the same testing script. Evaluate your test results.

Viewing your Tokenized Files

Viewing the tokenized files is another way to review the accuracy of your topic. You may want to view a tokenized file to see how the rules of the tokenization apply to your data. The tokenized data can be found in:

- `viaoice\topicfactory\us\[topic name]\trn.tok`

Look for the following things when reviewing the tokenized files:

- **Hyphenated Words**
Notice how the tokenizer parsed your hyphenated words. For example, the tokenizer might parse the word “semi-classic” with spaces before and after the hyphen, and it would appear in the `.tok` file as “semi - classic”. To correct this parsing error, add the word to your **Optional Special Words File**, reparse the data, and check the `.tok` file again to make sure the hyphenated word appears without spaces.

- **Abbreviations**

Notice how the tokenizer parsed your abbreviations. If the tokenizer interpreted a period in an abbreviation as a period at the end of a sentence, it put a space before and after the period. For example, “St. Louis” would be parsed “St . Louis”. To correct this parsing error, add the word to your **Optional Special Words File**, reparse the data, and check the .tok file again to make sure the abbreviated word appears without spaces.

- **Phrases**

In the tokenization process, phrases are surrounded by single quotes. When viewing the tokenized file, if a phrase, such as “New York” is not surrounded by single quotes, it has not been parsed as a phrase. If you view a phrase in your tokenized file parsed incorrectly with a space, add the phrase to your **Optional Special Words File**, reparse the data, and check the .tok file again to make sure the phrase appears with single quotes.

You may be required to train words that you added to your **Optional Special Words File**. After reparsing, check to see if the word or phrase appears in your **Not in Vocabulary** list box. If it does, add it to your **Topic Words**, and train if required.

Additional Pronunciations

In addition to editing baseforms for misrecognized words, you might discover when testing your topic that you need to add additional pronunciations for a word in your topic. For example, you've added the word "parmigiana" to your topic. When you trained the word you added the pronunciations "par mi zhah nuh" and "par mi jan uh". In testing your topic you discover the word is being misrecognized because it is also pronounced "par mi jan o". To add the additional pronunciation, refer to the Dictionary Builder helps.

Check and correct the baseforms for your new pronunciations. For more information on baseforms, see ["Create and Edit Baseforms" on page 26](#). Table 1 illustrates how the baseforms for these new pronunciations might look in Topic Factory:

Word Pronunciation	Topic Factory Baseform
pahr muh zhah nuh	P AA R M AX Z AA N AX
pahr muh jan a	P AA R M AX JH AA N AX
pahr muh jan o	P AA R M AX JH AA N OW

Table 1. Pronunciations and Baseforms for "parmigiana"

This process also works with a new topic word when the baseform is found in the backup dictionary. For example, you add the word "family" to your topic and the baseform found in the backup dictionary is "fam i lee". When testing your topic you find the word is being misrecognized because the tester pronounces the word "fam lee". Add the additional pronunciation to correct this problem. Table 2 illustrates what the Topic Factory baseforms might look like:

Word Pronunciation	Topic Factory Baseform
fam i lee	F AE M AX L IY
fam lee	F AE M L IY

Table 2. Pronunciations and Baseforms for "family"

Note:

Remember you can add up to five pronunciations for each new topic word. You can add pronunciations the first time you train your new words, or go back and add additional pronunciations as you discover the necessity through testing your topic.

Improving Topic Accuracy

To improve topic accuracy, continue the iterative process of:

- Testing and evaluating each generation of your topic
- Viewing the tokenized file
- Updating the valid words and reparsing your data
- Including more words in your topic for better coverage
- Adding supplemental source data for additional context
- Editing baseforms to correct pronunciation errors

Completing Your Topic

When you complete the testing process, and finish the Topic Factory build process for the last time, your files are ready to be packaged for installation. Your files are located in the following directory:

- `viaoice\topicfactory\us\[topic name]\release`

Now that your topic is complete, you're ready to update your "dummy" help file. It's all explained in ["Packaging Your Topic for Distribution" on page 37](#).

Packaging Your Topic for Distribution

The first time you built your topic, and every time you reparsed your data, you were asked to include the name of your help file. As suggested, you probably included a “dummy” file while you were still in the process of building your topic. Now your topic is complete! It’s time to document your topic for your users, and include your finished help file in your topic. This chapter describes how to package your topic for distribution.

Creating Your Help File

There are two ways to create your help file:

- Using any text editor, save it as an ASCII .txt file.
- Using any help authoring tool for Windows, create an .hlp file.

The purpose of your help document is to tell your users about your topic and how to use your topic. Here is an outline of suggested contents for your help file:

- What is a topic?
- What are the benefits of using your topic?
- Detailed information about your topic
- How to activate your topic
- Examples of specialized words in your topic and how to dictate those terms
- Where can the user get technical support for your topic?
- How can the user get information on additional topics from your company?

It’s important to test your document for usability with your topic, making sure all the dictation examples work properly. When your testing is finished, you’re ready to include your help file into your topic distribution package.

Integrating Your Help File

To integrate your final help file (.txt or .hlp) into your topic directory, copy the help file to your Topic Release directory, overwriting your “dummy” help file:

- `viaoice\topicfactory\us\[topic name]\release\[topic name].hlp` or `.txt`

Distributing Your Topic

It’s your choice to package your topic according to your distribution plans.

- If you ship on a CD, include all the files in the release directory
- If you distribute your topic from a Web site, you might choose to zip the files in your release directory and upload the zipped file to your Web site:

This appendix, intended for building a German-language topic, contains a few additional examples that take into account differences between U.S. English and German.

Gathering Data

As pointed out in the body of this guide, input data should be in ASCII (.txt) format. As there are a few characters in German, Umlaut, and scharfes S, which are often corrupted when transformed by import or export operations in common text editors. You have to make sure those characters are correct. If you do not, the list of unknown words produced when analyzing the text will be long and full of words split at those incorrectly transformed characters.

Tokenizing

In contrast to U.S. English, where “cross-country” can be split into "cross - country," German words containing a hyphen are not split. Thus, "Abteilungs-Chef" does not have to appear in your valid words file.

Lowercasing of Uppercase-Only Words

To prevent abbreviations like MIT from being interpreted as a misspelled version of the word "mit" and being transformed into "mit", put “MIT” into your valid words file.

Abbreviations

If you want your topic to support additional abbreviations, for example, to decide depending on context whether to write "z.B." or "zum Beispiel", you have to put these abbreviations into your valid-words file. Otherwise, sentence-boundary detection does not work.

Installation

All paths in the form `viaoice\topicfactory\us\` have to be changed to:

- `viaoice\topicfactory\gr\` for the German topic

Phoneme	Sound
AA	The "ah" sound in "are", "car", "far"-that is, followed by "r-like" sounds, but NOT as in "pot", "hot", "cot".
AE	The A sound, as in "add", "cat", or "pan".
AH	The stressed "uh" sound, as in "one" or "come", but NOT as in "obtain" or "alarm".
AO	The short O sound, as in "all", "off", or "caution", and also the "ah" sound in "pot", "hot", "cot".
AW	The "au" sound, as in "how", "about", "lout".
AX	The unstressed "uh" sound, as in "seven" or "accept", but NOT as in "enter".
AXR	The unstressed "er" sound, as in "another", "over", , but NOT as in "fur".
AY	The long I sound, as in "fire", "why", and "eye".
B	The B sound as in "be" or "able".
BD	The B sound at end of a word.
CH	The "ch" sound as in "cheap".
D	The D sound, as in "and" or "David".
DD	The D sound at end of a word.
DH	The voiced "th" sound, as in "the" or "either", but NOT as in "thesis" or "thing".
DX	The very short T or D sound, as in "butter", "greater", or "regarding".
EH	The stressed "eh" sound, as in "enter", "check", or "pleasant", but NOT as in "moment".
ER	The stressed "er" sound, as in "refer", "turn", or "her", but NOT as in "another".
EY	The long A sound, as in "became", "train", or "eight".

Phoneme	Sound
F	The F sound, as in "father" or "rough".
G	The hard G sound, as in "again" or "Peg".
GD	The G sound at the end of a word.
HH	The H sound, as in "here" or "who".
IH	The stressed short I sound, as in "it", "if", or "pick", but NOT as in "lasting".
IX	The unstressed short I sound, as in "discuss", "budgeted", "decided", or "saving", but NOT as in "it".
IY	The long E sound, as in "anyone", "obvious", "highly".
JH	The "j" sound, as in "jeep".
K	The K sound not at the beginning of a word, as in "because", "dark", or "scale".
KD	The K sound at the end of a word.
L	The L sound, as in "level" or "parallel".
M	The M sound, as in "am" or "must".
N	The N sound starting or the unaccented N in the middle of a word, as in "final", "not", or "none".
NG	The N followed by G or K sound, as in "bang", "think", or "singing".
OW	The long O sound, as in "quote", "open", or "go".
OY	The "oi" or long I sound, as in "avoid", "employ", or "boy".
P	The P sound not at the beginning of a word, as in "adoption", "amps", or "rapid".
PD	The P sound at the end of a word.
R	The R followed by a vowel sound in the same syllable, as in "abroad", "brace", or "read".
S	The S not followed by H sound, as in "sit", "circus", or "decide".
SH	The "sh" sound, as in "action", "shade", "beach", or "splash".
T	The T (not TH) sound not at the beginning of a word, as in "adapter", "retry", or "lets".
TD	The T sound at the end of a word.
TH	The unvoiced "th" sound, as in "thesis" or "thing", but NOT as in "the" or "either".
TS	The "t s" sound at the end of a word.

Phoneme	Sound
UH	The short U sound, as in "good", "put", "full", or "could".
UW	The stressed long U sound, as in "to", "use", "you", or "view".
V	The V sound, as in "eleven", "improve", or "very".
W	The W sound, as in "frequent", "question", "way", or "anywhere".
Y	The y sound that leads into a vowel, as in "emulate", "senior", "yes", or "you".
Z	The Z sound, as in "pans", "goes", or "zero".
ZH	The "zh" sound, as in "azure", "Asia", or "measure".

References in this publication to IBM products, programs, or services do not imply that IBM intends to make these available in all countries in which IBM operates. 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. Subject to IBM's valid intellectual property or other legally protectable rights, any functionally equivalent product, program, or service may be used instead of the IBM product, program, or service. The evaluation and verification of operation in conjunction with other products, except those expressly designated by IBM, are the responsibility of the user.

IBM may have patents or pending patent applications covering subject matter in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing
IBM Corporation
500 Columbus Avenue
Thornwood, NY 10594
U.S.A.

Asia-Pacific users can inquire, in writing, to the IBM Director of Intellectual Property and Licensing, IBM World Trade Asia Corporation, 2-31 Roppongi 3-chome, Minato-ku, Tokyo 106, Japan.

Licensees 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 Corporation, Department LZKS, 11400 Burnet Road, Austin, TX 78758 U.S.A. Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

Trademarks

The following terms are trademarks of the IBM Corporation in the United States or other countries or both:

- IBM
- Lotus
- VisualAge
- ViaVoice

The following terms are trademarks of other companies:

- Intel
- Intel Corporation

Microsoft, Windows, NT, the Windows 95 logo, Visual Basic, Visual C++, and Developer Studio are trademarks or registered trademarks of Microsoft Corporation.

Other company, product, and service names, which may be denoted by a double asterisk (**), may be trademarks or service marks of others.

Index

A

abbreviations, tokenizing, 17
accuracy testing, 28
activating a topic, 27
adding a new user, 15
adding pronunciations, 31
adding words in context, 13
adding words to a topic, 22
additional baseforms, 31
analyzing words, 24

B

backup dictionary, 31
baseforms, 31
 example of, 31
building a topic, 26
buttons
 Add, 18, 23
 Cancel, 23
 Context, 23
 Next, 18, 19
 Next context, 23

C

cache, 7
casing
 incorrect, 21
 tokenizing, 17
choosing topic words, 20
cleaning data, 14
compiling a topic, 26
completing your topic, 26
context, adding words, 13
context, viewing a word in, 23
conventions, used in this book, 6
coverage indicator, 23
create a topic, required skills, 9

creating a help file, 35
creating your help file, 35

D

data for testing, 14
describing a topic, 16
developer skills, 9
dictation vocabularies, 7
dictionary file, 17
distributing your topic, 35, 36
documenting your topic, 35
domain specific dictionary file, 17
domain, topic subject, 9
dummy file, 26, 33

E

editing baseforms, 31

F

fields
 Content, 23
file, special words, 17
frequency counts, 20
full-size vocabulary, 7

G

gathering data, 13
general dictation vocabulary, 7

H

hardware requirements, 11
help file
 creating, 35
 including in topic, 26, 35
 integrating, 36
hyphenated words, 17, 21

I

IBM ViaVoice Medical Vocabulary, 7
improving topic accuracy, 33
industry language model, 7
industry vocabulary, 7
installing Topic Factory, 11
installing your topic, 27
integrating your help file, 36
ISVs (Independent Software Vendors),
5

K

keep data for testing, 14

L

language model, 7
list of words, 16
listboxes
Not in Vocabulary, 20, 22, 23
Topic Words, 22, 23

N

naming a topic, 16

O

opening an existing topic, 18

P

packaging for installation, 33
packaging your topic, 35, 36
parsing, *see* tokenization rules, 16
Personal Language Model, 7
phrases
included in topic, 16
tokenizing, 17
pronunciation
phonetic representation, 10
when training new words, 10
pronunciations, 32

pronunciations, adding, 31

R

recognition engine data files, building,
26
recognition testing, 28
recording new words, 24
remaining words indicator, 23
reparsing data, 19
required skills, 9
restoring a topic, 18
reworking your topic, 28

S

scope of topic, 13
screens
Topic Factory--Baseform Editor, 31
Topic Factory--Complete Topic, 24
Topic Factory--Select Files, 18
selecting files, 18
selecting the base vocabulary, 15
selecting topic words, 20
skills, developer's, 9
software requirements, 11
source data
ASCII text format, 13
cleaning extraneous text, 13
data for testing, 14
gathering, 9, 13
multiple sources, 13
size, 13
special words file, 16
system requirements, 11

T

testing your document, 35
testing your topic, 28
tokenizing
abbreviations, 17
casing, 17

- dictionary file, 17
- hyphenated words, 17
- phrases, 17
- topic contents, 7
- Topic Factory, 7
 - cleaning data, 14
 - description of, 8
 - gathering data, 13
 - installing, 11
 - naming a topic, 16
 - select files screen, 18
 - special words file, 16
- Topics
 - definition of, 7
 - describing, 16
 - naming, 16
- topics, what is a topic, 7
- training new words, 24

U

- uninstalling Topic Factory, 12

V

- VARs (Value Added Resellers), 5
- viewing a word in context, 23
- Vocabulary and Topic Installer, 27

W

- Web, distributing your topic on, 36
- what is a topic?, 7
- what is Topic Factory?, 8
- wizard, Topic Factory, 8, 9
- words file, special, 16
- working with baseforms, 31

