



NDEV Mobile Developer Program
Guide to Custom Vocabularies

1.5

Notice

Guide to Custom Vocabularies

Copyright © 2013, 2014, 2015 Nuance Communications, Inc. All rights reserved.

Published by Nuance Communications, Inc.

One Wayside Road, Burlington, Massachusetts 01803, U.S.A.

Last updated February 6, 2015.

Nuance Communications, Inc. provides this document without representation or warranty of any kind. The information in this document is subject to change without notice and does not represent a commitment by Nuance Communications, Inc. The software and/or databases described in this document are furnished under a license agreement and may be used or copied only in accordance with the terms of such license agreement. Without limiting the rights under copyright reserved herein, and except as permitted by such license agreement, no part of this document may be reproduced or transmitted in any form or by any means, including, without limitation, electronic, mechanical, photocopying, recording, or otherwise, or transferred to information storage and retrieval systems, without the prior written permission of Nuance Communications, Inc.

Nuance and the Nuance logo are trademarks or registered trademarks of Nuance Communications, Inc. or its affiliates in the United States and/or other countries. All other trademarks referenced herein are the property of their respective owners.

Table of Contents

1	Overview	2
2	Guidance for Building a Vocabulary	2
3	Limitations	3
4	Custom Vocabulary Format of Phrase Lists	4
5	Using Application Vocabularies	4
6	Application Vocabulary Management Process	5

1 Overview

NDEV Speech recognition uses language models to recognize speech utterances. Language models consist of hundreds of thousands of words, plus estimates of the likelihood of words occurring in the various word sequences. NDEV supports several language models such as a general dictation model, a web search language model and a TV language model. While these language models are designed to serve the needs of a wide variety of applications and customers there is often the need to modify the words or recognition probabilities for a particular recognition context.

The Custom Application Vocabularies feature lets you create and upload application-specific vocabularies to the Nuance Cloud Service via the NDEV portal. When a speech recognition request is made, the custom vocabulary augments the particular language model associated with the vocabulary. The custom vocabulary increases the efficiency and accuracy of your application in two ways: by increasing the probability that a spoken word or phrase is recognized accurately; and by increasing the probability that the correct written representation of the utterance is displayed. For example, you can upload a vocabulary for music and song titles so that the system returns “Eminem” with higher probability than “M&M.”

Application vocabularies are useful for words and phrases whose method of recognition is the same for all users of an application (for example, product names) as opposed to a particular user.

Application vocabularies are supported for all speech recognition languages.

It is important to note that custom vocabularies are different than the constrained speech recognition grammars many developers are familiar with. Using constrained grammars results in high accuracy of words that are in the grammar and low (or no) recognition of words or phrases that are not in the grammar. With custom vocabularies our large language models are still used and the vocabulary simply adjusts the recognition probabilities of that large model. As a result, using a vocabulary will not change your results as much as a conventional grammar would. For example, you can expect to still get word results that are not in the vocabulary or “out of grammar.” The grammar recognition mode feature makes the vocabulary act much more like a traditional recognition grammar but as we are still using the underlying language model even then you may get “out of grammar” results.

2 Guidance for Building a Vocabulary

- Repetitions: Repeating a phrases will not increase recognition probability of the phrase
- Context: If known, it is helpful to include phrases you expect will be spoken for the intended use-case. They do not have to cover the entire expected utterance, but should have some

context around the key words of interest. For example, if the word “CoolPix” must be added, and it would typically be said in the context of “search for CoolPix camera” then adding the entire phrase will reduce the error rate. Adding similar phrases (which may include different contextual words or none at all) should also see reduced error rate but to a possibly to a lesser degree. The feature is also useful to resolve homonyms. For example, if needing to favor the word “cereal” instead of “serial”, then the phrase “cereal” or “breakfast cereal” should be added.

- **Text Cleanliness:** The impact of custom vocabularies on speech recognition accuracy depends on the quality of the vocabulary text and every effort should be made to clean the text from irregularities while preserving the integrity of the original text:
 - The text should use consistent UTF-8 encoding. Accents and casing should be preserved (e.g. Café Romi).
 - The text should be plain text without any font formatting.
 - The text should be formatted like it was in a newspaper, i.e. it should have proper spacing and capitalization.
 - In Asian languages, latin text (A-Z), and numbers (0-9) should be provided in half-width instead of full-width character representation
 - An utterance refers to an example of what the user is expected to say to the system. However, it can also be a fragment (e.g. a named entity like movie name, point-of-interest).

3 Limitations

Programmatic Limits

The following limits on vocabularies will vary by tier. The specific limitations for your tier are described in the pricing section of the NDEV portal

- **Data Item Limit:** Each phrase in a vocabulary is referred to as a data item.
- **Word Count Limit:** The total number of words in a vocabulary.
- **Vocabulary Files Size Limit:** The size of a vocabulary file in KB.
- **Total Vocabulary Limit:** The total number of vocabularies you can store on our servers (for development and production purposes.) Vocabularies are listed in the Manage Vocabulary Files section of the Manage My Vocabularies Page. Note, different versions of a vocabulary count toward this limit.

Performance Limitations

Using custom vocabularies will add some delay / latency to recognition requests. As the vocabulary needs to be loaded for every recognition request this latency occurs for every recognition request. The amount of added latency should be negligible for the vast majority of self-serve customers. Platinum customers, however, should exercise more care with respect to latency given their ability to upload much larger vocabularies. Particular care should be exercised when working with vocabularies larger than 2,000 data items. The key variables affecting latency in order of impact are:

- **Number of Out of Vocabulary (OOV) Words:** This is the largest source of latency. The more

words in the vocabulary that are NOT in the underlying topic / language model you are using the greater latency end-users will experience. Our underlying languages models are very large (containing hundreds of thousands of words) so this should not happen frequently but can occur when using very obscure words and particularly when using words that are in a different language. We do not at this time have a tool that you can use to determine the number of OOV words in your vocabulary.

- Number of data items: A data item is a single row (which can be a single word or phrases) in your vocabulary.
- Number of words: The total number of words in your vocabulary.

Other Limitations

- The results mapping feature described below requires customer support assistance to activate for the vocabularies that use it.

4 Custom Vocabulary Format of Phrase Lists

General Format

Content is in UTF-8 format

```
<phrases>  
    <phrase>custom phrase 1</phrase>  
    <phrase>custom phrase 2</phrase>  
    <phrase>...</phrase>  
</phrases>
```

5 Using Application Vocabularies

Application vocabularies must be uploaded through the NDEV web portal. They must be associated with a particular language, language model (topic), and Application ID (development or production). The vocabulary will automatically affect the recognition results of all recognition requests made for the particular Application ID, language, and topic associated with that vocabulary. You will also assign the vocabulary a “weight” which reflects the importance of the vocabulary items relative to the underlying language model. The default weight is 80% and can be anywhere between 10% and as 90%. Note, that even when using an application vocabulary and even with a weighting of 100%, the underlying language model assigned to the vocabulary is still being utilized so that the words in the language model are still able to be recognized.

Grammar Recognition Mode

Grammar recognition mode seeks to restrict recognition to only the words or phrases in your vocabulary. This greatly increases the recognition accuracy of the words in your vocabulary and is particularly useful when you can anticipate the words end-users will say in advance. To use grammar recognition mode you will need to indicate that you wish this feature to be used when

associating a vocabulary file to a development or production App ID. Please note that grammar recognition mode does not perfectly restrict recognition to words in the vocabulary. You may occasionally receive results of words that are not in the vocabulary.

Important

For the application vocabulary to be used for recognition you must use the 6 character languages codes (XXX-XXX) available on the NDEV web portal when making recognition requests.

6 Application Vocabulary Management Process

Your application vocabularies will proceed through a three-step process.

1. **Upload** – The first step is to upload your Vocabulary to our servers. When you upload the vocabulary you will also assign a language to the vocabulary. After uploading your vocabulary you will see the vocabulary in the “Manage Vocabulary Files” Section of the “Manage My Vocabularies” page.

2. Test / Develop

- In order to use your vocabulary you next need to **associate** it with your sandbox Application ID and enable it (enabled by default when you associate with your development App ID.) When you **associate** the vocabulary with your sandbox App ID you will also specify what topic (Dictation, WebSearch, DTV) your vocabulary will be used with, whether you wish to use grammar recognition mode, and the weight you want assigned to it for recognition requests if you are not using grammar recognition mode.
- After you have **associated** the vocabulary you will see you vocabulary in the Development Vocabulary Assignments section of the page. If the status shows “Pending” give it a few minutes until it shows “live.”
- When the vocabulary is live it will be used for all speech recognition requests for the development App ID made for the language of the vocabulary and the topic you specified. Note: You **MUST** move your vocabulary into this stage before you can use your vocabulary with a production application.
- From the Development Associations section of the Manage My Vocabularies page you will be able to adjust meta-information for the development App ID association such as the topic, whether grammar recognition mode is used, and the weight for the vocabulary.
- You will also be able to disable the vocabulary without removing the association or you can remove the association altogether. Note, removing the association does not delete the vocabulary. To do that you must delete it from the Manage Vocabulary Files section of the Manage My Vocabularies page. Currently you can only use a single vocabulary for a particular language and topic. Disabling and enabling the vocabulary (without removing the association) lets you easily test multiple vocabularies with your single development application ID.

3. Production

- After you have completed testing your vocabulary you will need to **approve** your vocabulary for production and **associate** it with a particular production application ID. The easiest way to do this is to “Promote” a Vocabulary from the Development Associations section of the Manage My Vocabularies page. When you do this the vocabulary will automatically be “approved” and you will be presented with a page where you can associate it with one or more Production Application IDs. When making the associations you will re-iterate the topic your vocabulary will be used for, whether you want to use grammar recognition mode and the weight you want assigned to it if you are not using grammar recognition mode.
- After you have made an association you will see the association in the Production Vocabulary Associations section of the Manage My Vocabularies page. When the status indicates Live (which should take less than 5 minutes) your vocabulary will be used for all recognition requests for the particular production Application IDs, language of the vocabulary and topic you have assigned.
- At this point you will be able to adjust meta-information production Application ID association such as topic, grammar recognition mode, and weight for the vocabulary. When editing an association you can also directly apply a vocabulary file which you have already approved for production. You can even copy the assignment to a different application ID. This allows you to easily use the same vocabulary file and parameters with a different application.
- And as with your development App ID association you can now also disable/enable the vocabulary as well as completely remove the association for a particular Application ID.
- You can delete vocabularies from the Manage Vocabulary Files section of the page. But in order to do this you will first need to remove all development and production vocabulary associations that may exist.