Google Search Appliance

Search Appliance Internationalization

Google Search Appliance software version 7.2 and later

Google, Inc.
1600 Amphitheatre Parkway
Mountain View, CA 94043
www.google.com

GSA-INTL_200.01
March 2015

© Copyright 2015 Google, Inc. All rights reserved.

Google and the Google logo are, registered trademarks or service marks of Google, Inc. All other trademarks are the property of their respective owners.

Use of any Google solution is governed by the license agreement included in your original contract. Any intellectual property rights relating to the Google services are and shall remain the exclusive property of Google, Inc. and/or its subsidiaries ("Google"). You may not attempt to decipher, decompile, or develop source code for any Google product or service offering, or knowingly allow others to do so.

Google documentation may not be sold, resold, licensed or sublicensed and may not be transferred without the prior written consent of Google. Your right to copy this manual is limited by copyright law. Making copies, adaptations, or compilation works, without prior written authorization of Google. is prohibited by law and constitutes a punishable violation of the law. No part of this manual may be reproduced in whole or in part without the express written consent of Google. Copyright © by Google, Inc.

Contents

Search Appliance Internationalization

About this Document

This document covers all aspects of how the search appliance handles different languages and character encodings for search, indexing, serving, and administration.

Which Languages are Searchable and Indexable?

The search appliance can index and search languages whose writing systems can be expressed in the UTF-8 encoding. This includes European and other languages that use the Latin alphabet, Chinese, Japanese, Korean, Arabic, Hebrew, and Thai. Search result quality may vary among languages, depending on the language resources, such as synonym files, that are available on the search appliance.

To improve search quality for your language, upload synonym files on the Query Settings page. If the search appliance includes a synonym file for your language, you can improve search quality by providing synonyms for your business's internal abbreviations, code names, and other terms particular to the business. For more information on synonym files, see "Best Practices" in *Creating the Search Experience*.

Which Character Encodings can be Crawled and Indexed?

The search appliance can crawl and index documents in all common character encodings. If an unusual encoding is found in your document corpus, Google recommends that you use a development appliance to test the encoding.

What is the Recommended Character Encoding for Documents?

Google strongly recommends that whenever possible, you encode documents to be crawled or fed using the UTF-8 character encoding. If your documents use other character encodings, such as legacy systems or documents that use national character set encodings, those documents can be crawled or fed.

How are Bi-Directional Languages Handled?

Hebrew and Arabic are bi-directional languages, meaning that text in those languages is written and read from right to left. Documents in Hebrew and Arabic can be indexed and searched. Google provides query expansion and synonym data for Hebrew and Arabic or you can provide your own synonym files.

How are Chinese, Japanese, and Korean Handled?

Documents in Chinese, Japanese, and Korean can be searched and indexed. Google does not provide query expansion or synonym data for these languages. However, you can provide your own synonym files.

All regional Chinese dialects are written with Traditional Chinese or Simplified Chinese characters. The Google Search Appliance can crawl, index, and search documents written in both Traditional and Simplified Chinese characters.

How is Segmentation Used?

The search appliance uses segmentation to determine word boundaries in text that is being indexed. For example, in the sentence "The dog was sleeping," the search appliance uses the spaces between the words to determine word boundaries. Some languages, including Chinese, Japanese, and Korean, do not use spaces as boundaries between words. When the search appliance crawls and indexes languages that do not use spaces between words, it uses an internal dictionary to determine where to segment the text into words.

How Does the Search Appliance Detect the Language of a Document?

The search appliance's software analyzes the content of a document, including character and word frequency, to determine the document's language. If the web page or other file does not contain sufficient text to determine the language, the search appliance uses the Content-Language header in the HTTP response for the page. If the content-type header or http-equiv meta tag for the web page or file specifies a character encoding, that encoding overrides language detection based on the page content. For example, if a document's HTTP headers specify the GB2312 encoding, the search appliance recognizes the page as Simplified Chinese, regardless of the page content.

Which Character Encodings can be used to Enter Queries and Serve Results?

To support searching documents in multiple languages and character encodings, Google provides the `ie` and `oe` parameters:

- The `ie` parameter indicates how to interpret characters in the search request.
- The `oe` parameter indicates how to encode characters in the search results.

To appropriately decode the search query and correctly encode the search results, supply the correct `ie` and `oe` parameters, respectively, in the search request. When you are providing search for multiple languages, Google recommends using the utf8 encoding value for the `ie` and `oe` parameters. To view a list of supported character encodings, see the "Internationalization" section of the *Search Protocol Reference*.

How Does the Search Appliance Detect the Language of a Query?

The search appliance determines the language of a query by taking into account the words in the query, the front end in use for the query, and the search user's language setting in the browser.

In What Languages are the Admin Console and Help System Displayed?

This section lists the languages in which the Admin Console and help system are displayed.

The Admin Console and online Help system are available on the search appliance in the following languages:

Arabic	English (UK)	Hungarian	Russian
Basque	English (US)	Italian	Slovak
Catalan	Finnish	Japanese	Spanish
Chinese (Simplified)	French	Korean	Swedish
Chinese (Traditional)	Galician	Norwegian	Thai
Czech	German	Polish	Turkish
Danish	Greek	Portuguese (BR)	Vietnamese
Dutch	Hebrew	Portuguese (PT)	

How Does the Search Appliance Determine the Display Language for the Admin Console and Help System?

The language setting in a browser determines the language in which the search appliance displays the Admin Console and help system.

When the browser makes a request to the search appliance, the request includes one or more preferred languages. The Admin Console or requested help page is displayed in the first preferred language in which the search appliance page is available.

For example, if the preference list contains Armenian, French, and English, in that order, the Admin Console and help pages are displayed in French, the first language in which the pages are available. If the order is changed to English, French, and Armenian, the pages are displayed in English.

If the preference list does not contain a language in which the Admin Console and help system can be displayed, the search appliance defaults to English. For example, if the preference list contains Arabic and Gaelic, the display language is English.

For information on changing language settings, refer to the help system for your browser.

What Languages and Character Sets can be Typed into the Admin Console?

The default character encoding in modern browsers is UTF-8, a Unicode encoding capable of displaying characters in most writing system. (For a complete listing of the writing systems that can be displayed with Unicode encodings, see the "Unicode Character Code Charts By Script," at http://www.unicode.org/charts/.) The UTF-8 encoding supports a full range of characters in any language.

There are some restrictions on collection names and other search appliance parameters and fields. For more information on these parameters, see "Are there Restrictions on Using Non-ASCII Characters in the Admin Console Fields?" on page 7.

Are there Restrictions on Using Non-ASCII Characters in the Admin Console Fields?

You can use international characters in an Admin Console field unless the documentation for a particular field or feature states otherwise.

The following fields are restricted to alphanumeric ASCII characters, underscores, and hyphens only:

- Collection names
- The name you assign on the Admin Console to a Query Expansion synonym file. This is not the name of the file to which you browse in the **File** field, which may contain any UTF-8 characters, but the name you assign in the **Search > Search Features > Query Settings > Name** field.
- The name you assign on the Admin Console to a Query Expansion blacklist file. This is not the name of the file to which you browse in the **File** field, which may contain any UTF-8 characters, but the name you assign in the **Search > Search Features > Query Settings > Name** field.

Can Searches be Made Accent-Insensitive?

By default, searches on the Google Search Appliance are accent-sensitive. For example, if you search for the term `distribuicao`, it will not match the Portuguese word `distribuição`. Accent-sensitivity works for most commonly-occurring words that contain diacritical marks in each language.

There are two ways to make searches accent-insensitive: by using the `lr` parameter or by using the Accept-Language HTTP header.

Most of the time, accent-insensitive search works in both directions. For example, when accent-insensitive search is enabled, searching for `distribuicao` returns results containing `distribuição`, and searching for `distribuição` returns results containing `distribuicao`.

You use the `lr` parameter in search URLs. For example, if you set the `lr` parameter to the value `lang_pt`, for Portuguese (`lr=lang_pt`), a search for `distribuicao` will match `distribuição`.

Accent-insensitive search with the `lr` parameter is available in German, French, Spanish, Finnish, Norwegian, Swedish, and Portuguese. For more information on how to use the `lr` parameter in constructing search URLs, see the *Search Protocol Reference*.

Use the Accept-Language HTTP header to enable accent-insensitive search for the most common words containing diacritical marks in the specified language. For example, `Accept-Language: pt` in the request results in case-insensitive search in Portuguese. the Accept-Language header is set in end-user browsers. Refer to the documentation for your browser for more information.

Take note that accent-insensitive search works only for languages that are part of a language bundle that is currently enabled. For more information about language bundles, see “How do Language Bundles Work?” on page 11.

Which Language-Related Settings are User-Configurable?

Search users can configure the preferred language setting in their browsers.

Administrators can modify the XSLT style sheet to control various language-related features. For example, if you use older web servers, you can submit queries to and receive results from the search appliance in a National Character Set rather than UTF-8. For more information, see the “Internationalization” and “Language Filters” sections of the *Search Protocol Reference*.

Administrators can configure the following front-end settings and features:

- Query expansion, a feature that allows you to enable use of preinstalled synonym files or upload your own synonym and blacklist files
- Style sheets, which allow you customize the languages in which the search page, advanced search, and results pages are presented

- Language filtering, which allows you to restrict the languages of documents returned in searches. Language filtering, which is enabled in front ends, is available in the following languages:

Arabic	Estonian	Icelandic	Polish
Chinese (Simplified)	Finnish	Italian	Portuguese
Chinese (Traditional)	French	Japanese	Romanian
Czech	German	Korean	Russian
Danish	Greek	Latvian	Spanish
Dutch	Hebrew	Lithuanian	Swedish
English	Hungarian	Norwegian	Turkish

For more information on query expansion, front ends, and language filtering, see *Creating the Search Experience*.

Which Languages are Spelling-Checked?

Spelling check at serve time is available for English, Portuguese, French, Italian, German, Spanish, and Dutch. As new documents in the supported language are added to the document corpus and indexed, the search appliance learns new words and is able to provide spelling check for those new words.

How Does the Search Appliance Make Spelling Suggestions?

The search appliance makes spelling suggestions in response to search queries. The following conditions must be met before the search appliance makes spelling suggestions:

- The search appliance spelling database must contain an entry that matches the search term or the search appliance must have learned the correct spelling from the corpus of documents it crawls.
- The language of the search query must be supported by the active language bundle that is installed on the search appliance.

  The default language bundle supports English, Portuguese, French, Italian, German, Spanish, and Dutch. The appliance typically detects the language of a query from the two-letter country code in the Accept-Language header sent by the web browser. Browser documentation provides instructions on how to change the browser language.

In some cases, the query needs additional context. For example, a one-word search term may not trigger query rewrite, but adding more terms to the query might result in a spelling suggestion for the original word in the query.

The default language bundle that is installed on the search appliance has some additional rules governing spelling suggestions:

- If the browser language is English (en), Portuguese (pt), French (fr), Italian (it), German (de), or Spanish (es), the search appliance returns spelling corrections. This also applies to variants such as pt-PT and pt-BR.
- If the browser language is Japanese (ja), traditional Chinese (zh-TW), Korean (ko), or simplified Chinese (zh-CN), the search appliance returns English spelling corrections.
- If the browser language is any other language, the search appliance does not return spelling corrections.

The appliance tries to learn new words from the corpus of crawled documents that are in the languages of the active language bundle. The search appliance is then able to provide spelling suggestions.

Which Languages can use Query Expansion?

Query expansion is available by default for Arabic, Czech, Dutch, English, French, German, Italian, Polish, Portuguese, Russian, Slovak, Spanish, and Swedish, and can be made available in other languages if a language bundle containing those languages is installed and active. Query expansion adds extra terms to the search query.

There are three types of query expansion. In the first type, diacritical query expansion, the search appliance automatically performs diacritical query expansion for search queries in any language, by adding accented and non-accented variants of terms to the search query.

In the second type, contextual query expansion, the search appliance adds synonyms for terms to the search query. For example, the search term “latest apple” might be expanded to include “apples,” “fruit,” and “ipod.” The search appliance performs this type of query expansion only when the following conditions are met:

- The appliance query expansion database must contain an entry that matches the search term.
- The language of the query is supported by the active language bundle that is installed on the appliance. The default language bundle supports English, Portuguese, French, Italian, German, Spanish, and Dutch.

In some cases, the query might need additional context. For example, a single-word search term might not trigger query expansion, but adding additional search terms results in expansion of the original word in the query.

In the third type of query expansion, non-contextual query expansion, the query term is replaced. For example, the term “apple” would be expanded to “apple, apples.” Non-contextual query expansion is available only for the languages in the default language bundle, which are English, Portuguese, French, Italian, German, Spanish, and Dutch.

Search appliance users do not see the effects of query expansion on the original search terms.

You can also create a local query expansion policy using preinstalled and custom synonyms and blacklist files for languages that use the Latin-1 (ISO_8859-1) alphabet, provided that files containing accented characters are UTF-8 encoded. For more information on query expansion, see “Using Query Expansion to Widen Searches” *in Creating the Search Experience*.

Which Languages can use Dynamic Result Clustering?

Dynamic result clustering is a feature that presents a search user with a list of search terms that can be used to narrow a search. For example, with dynamic result clustering enabled, a user searching on the term opera might see the following result cluster:

narrow your search

opera history
opera singers
opera companies
opera composers
opera houses

Dynamic result clustering is enabled in individual front ends. The feature works with most languages, but is most effective when the results contain sufficient documents to allow the search appliance to categorize them into topics. If a crawled corpus contains a small number of documents in a particular language, dynamic result clustering will be less useful in that language.

For more information on dynamic result clusters, see "Best Practices" in *Creating the Search Experience*.

How do Language Bundles Work?

Language bundles provide resources that are used for spelling support and query expansion. Google provides a default, preinstalled, language bundle for English, Portuguese, French, Italian, German, Spanish, and Dutch.

Only one language bundle can be active on a search appliance. For example, if you install a language bundle for Hungarian, Russian, and Finnish and make query expansion active in a particular front end, the default language bundle is no longer active.

The following table lists the language bundles that are currently available, with links to bundles other than the default language bundle. The table will be updated as new bundles become available.

Language Bundle Name	Languages
Default	English Portuguese French Italian German Spanish Dutch
All languages (http://dl.google.com/dl/enterprise/all_langs-lang-pack-2.2-1.bin)	Arabic Danish German Greek English Spanish Finnish French Hungarian Italian Hebrew Japanese Korean Dutch Norwegian (includes both Bokmål and Nynorsk) Polish Portuguese Romanian Russian Swedish Thai Turkish
Northern European (http://dl.google.com/dl/enterprise/mea_north_eastern-lang-pack-2.1-1.bin)	English German Dutch Swedish Norwegian (Bokmål) Norwegian (Nynorsk) Danish Finnish Russian Finnish Spanish
Scandinavian (http://dl.google.com/enterprise/scandinavia-lang-pack-2.1-1)	English German Dutch Swedish Norwegian (Bokmål) Norwegian (Nynorsk) Danish Finnish

Language Bundle Name	Languages
Middle Eastern (http://dl.google.com/dl/enterprise/mid_east-lang-pack-2.1-1)	Arabic Greek English French Hebrew Turkish
Eastern European (http://dl.google.com/dl/enterprise/east_europe-lang-pack-2.1-1)	German English French Hungarian Hebrew Polish Romanian Russian
egfisdfp (http://dl.google.com/enterprise/egfisdfp-lang-pack-2.1-1.bin)	English German French Italian Spanish Dutch Finnish Polish

For more information on language bundles, see "Changing Languages for Query Expansion and Spelling Suggestions" in *Creating the Search Experience*.