
Linguistic Features for Primo VE

This section describes the various linguistic features that Primo supports.

Language Detection

In order to offer language-based services, Primo must first detect the language of the indexed text and the query. Currently, Primo can detect the following languages:

- Latin-based: English, Spanish, Italian, German, French, and Danish.
- Asian: Chinese, Japanese, and Korean. If the character is Chinese and the locale of Primo is Japanese or Korean, Primo uses the locale of the selected language.
- Other languages that have a specific character range: Hebrew, Arabic, and so forth.

Language detection is based on comparing the words of the record and the query with a dictionary. If fifty percent or more of the words match, the language is identified.

Stop Words

Stop words are included in phrase searches and omitted from keyword searches. For example, if a user searches for **the adventures of huckleberry finn**, Primo performs the following searches:

```
the adventures of huckleberry finn
```

or

```
adventures huckleberry finn
```

Primo uses stop word lists during indexing and searching.

Author Names

Primo treats words with O' apostrophe as a stop word in many Latin languages and indexes them as two separate words. This happens also for authors such as O'Leary, which is indexed as **o** and **leary**. As a result, a search for **Oleary** will not retrieve the same number of results as **O'Leary**. When users search for names that typically include apostrophes but do not include the apostrophe, Primo will also search for the name as if the users had included the apostrophe. For example, if the user's query is **Oleary**, Primo will change the query to search for **oleary** or **o leary**.

Stemming

Stemming is a process that reduces inflected (or sometimes derived) words to their stem, base, or root form. Primo uses stemming when a search returns fewer results than 25. If the search returns fewer results than this amount, Primo stems the search terms using the Kstem stemmer.

Primo identifies the language of the query and applies relevant stemming logic which can be different based on the language. In cases where Primo cannot identify the language of the query terms it will look at the user's interface language to define the language logic to apply.

Note

Primo does not unstem terms with the exception of pluralizations. If the result set is lower than the default threshold, Primo will pluralize terms, ranking their results lower. For example, a search for **wild flower** expands to **wild AND (flower OR flowers^{0.5})**.

The following parameter on the Discovery Customer Settings page ([Configuration Menu > Discovery > Other > Customer Settings](#)) allows you to disable or limit the use of stemming:

- **maximum_results_for_stemming** – When set to **true**, this parameter sets the maximum number of results returned before the system uses stemming to return more results. If this parameter is set to 0, stemming is not used to return results. By default, this parameter is set to 25 results.

Synonyms

Primo adds the following types of synonyms to a search query:

- Numbers – when a search contains a digit, Primo adds the spelled out number to the search query. For example, Primo adds the word **ninth** to a search query for **9th**.
- US or British spelling – when a search contains a word spelled according to US or British spelling, Primo adds the corresponding synonym to the search query. For example, Primo adds the word **colour** to a search query for **color**.
- Commonly misspelled words – for commonly misspelled words, Primo adds the word spelled correctly to the search query.

In addition to the synonym, Primo includes the original search term in the query. For example, if the query is **fifth dimension**, Primo searches for **(fifth OR 5th) AND dimension**.

Primo applies a different set of Synonyms lists based on the language recognition.

The following parameter on the Discovery Customer Settings page ([Configuration Menu > Discovery > Other > Customer Settings](#)) allows you to disable the use of synonyms:

- **disable_synonyms** – When set to **true**, this parameter disables the use of synonyms in search queries. By default, this parameter is set to **false**.