LanguageWare Resource Workbench 7.2

Custom dictionaries

Introduction

- **Module overview**
  - How to create and configure Custom Dictionaries.
  - Best practices
- **Target audience:**
  - All audiences
- **Prerequisites:**
  - Install LanguageWare® Resource Workbench (LRW)
  - Create a new project
- **Version release date:** LRW 7.2, ICA 2.2, released October, 2010

Module objectives

After this module you will be able to:

- Create and customize a custom dictionary
- Add entries to a custom dictionary
- Import and export data from a custom dictionary
- Use a custom dictionary to annotate a document
- Use features to add useful information to the entries in the custom dictionary

Module roadmap

- **Custom dictionary database**

  What is it?

  How to configure it?

  How to add, edit, export and import data?

- **Summary and best practices**
- **Sample exercises**

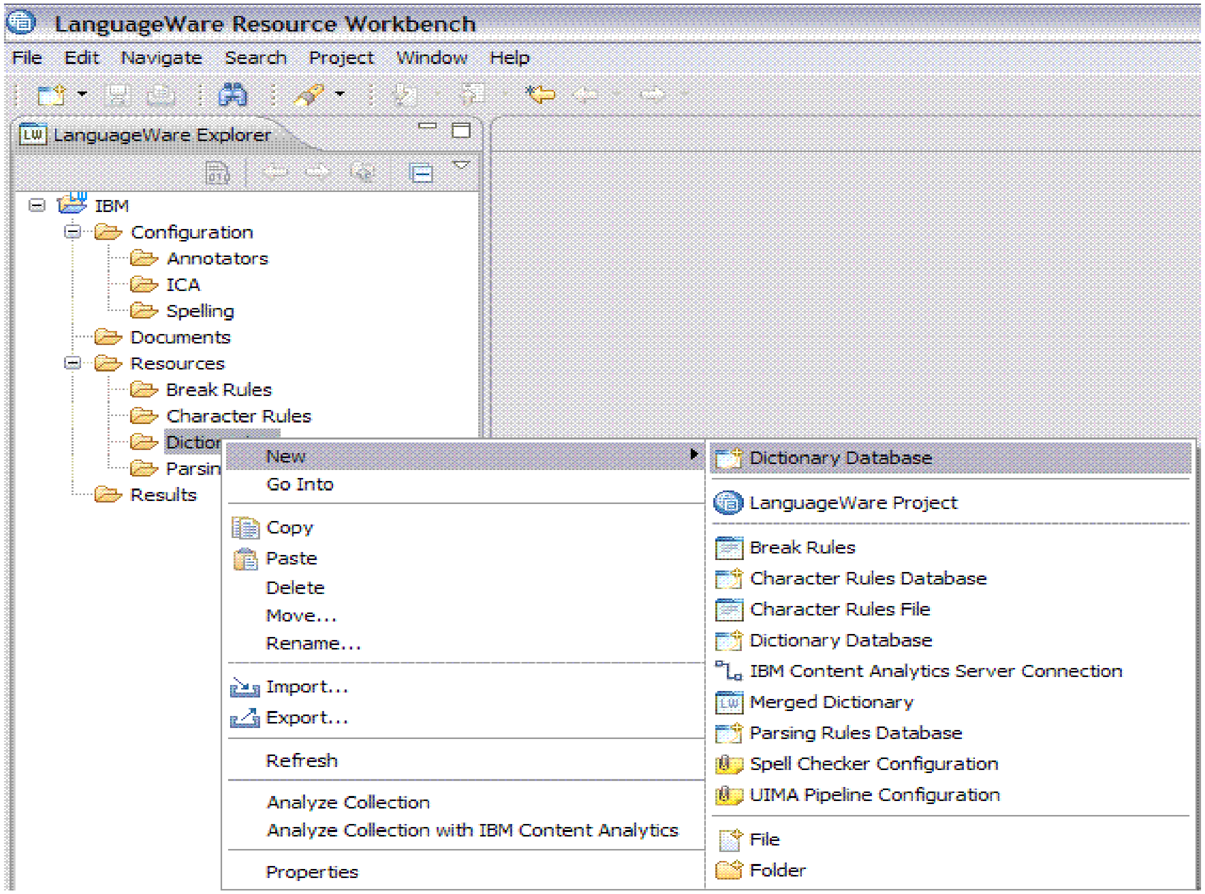
Custom dictionary
What is it?

- **General**
  - – *List of words/phrases to be searched in texts*
- **Specific**
  - – *Type one: List of search entities (place names, first names, Drug names, IBM Products...)*
  - – *Type two: Triggers or indicators that occur in the proximity of interesting data (company indicators, eg. “& Co”, address indicators, eg “st.”,”ave.”...)*

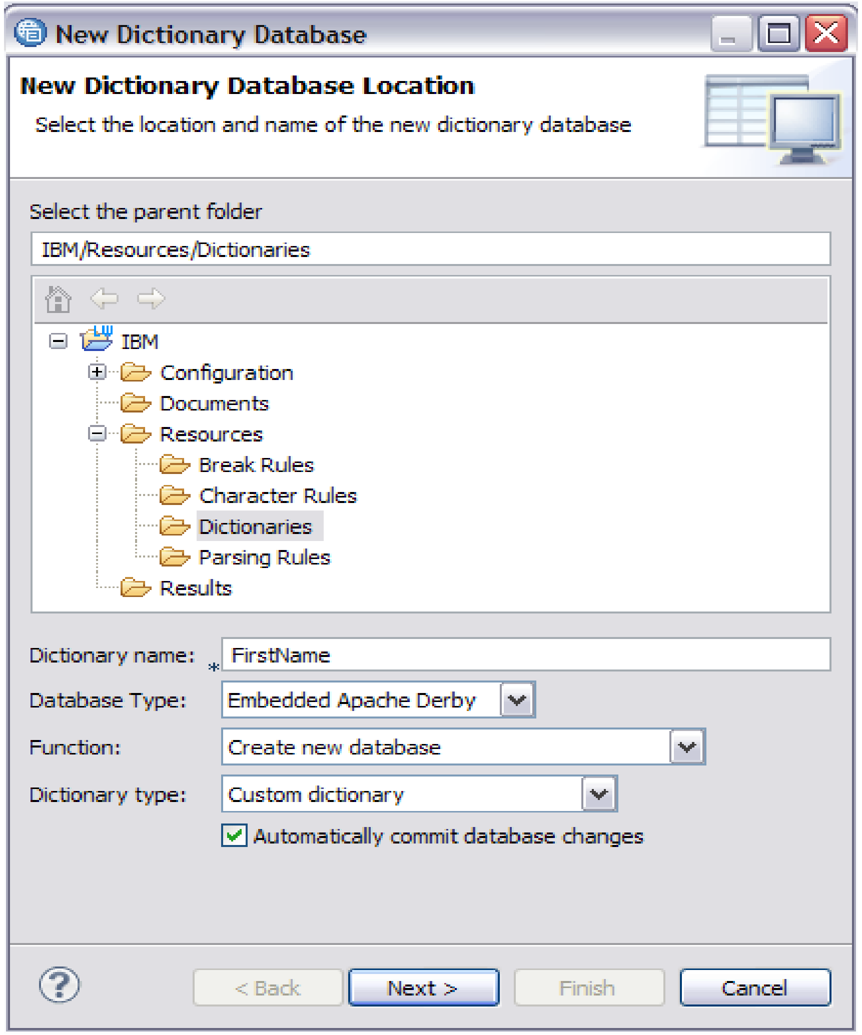
Custom Dictionary

How to create and configure it?

- *right click on the Dictionary folder in the project, select New/Dictionary Database*



- Specify the parent folder and the name of the dictionary (names should be representative of the type of entries of the dictionary). Then click Next.



New Dictionary Database

New Dictionary Database Location

Select the location and name of the new dictionary database

Select the parent folder

IBM/Resources/Dictionaries

Dictionary name: * FirstName

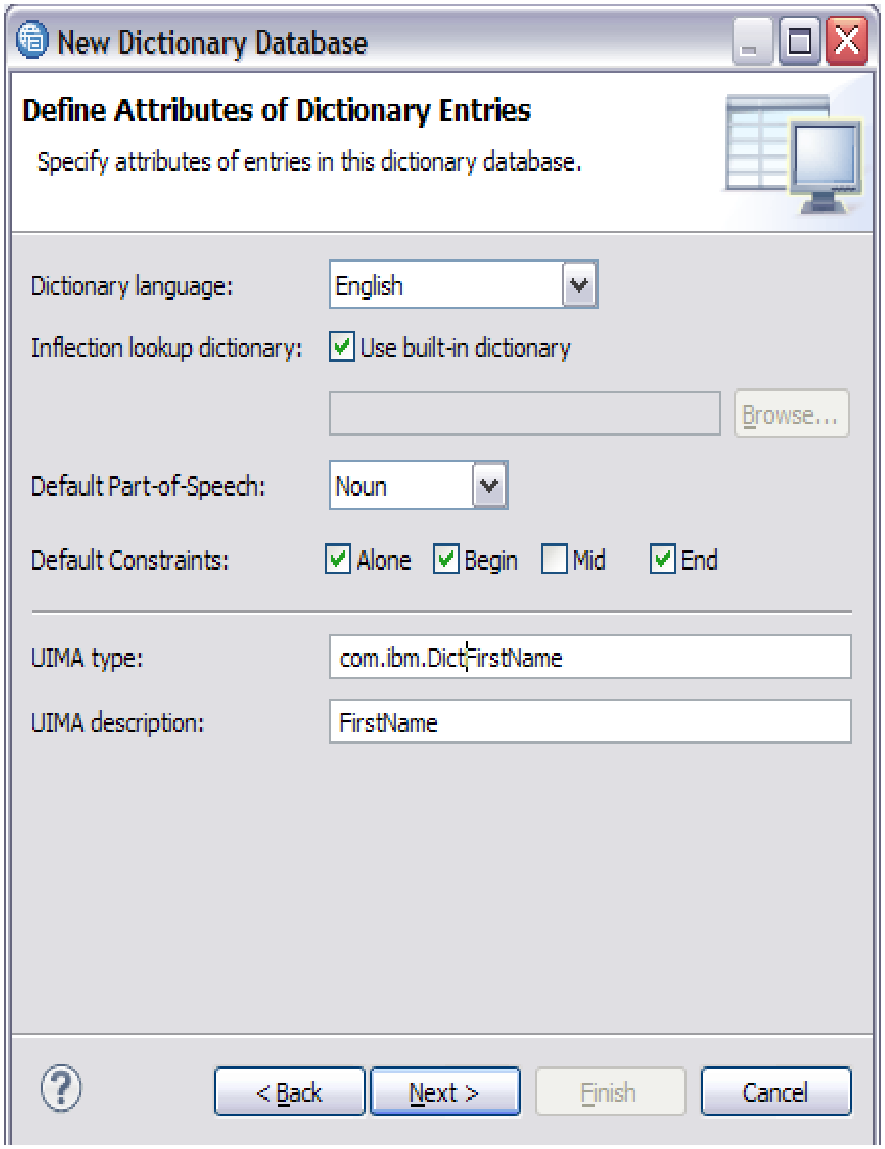
Database Type: Embedded Apache Derby

Function: Create new database

Dictionary type: Custom dictionary

☑ Automatically commit database changes

- Define attributes of the dictionary entries:
  - **Dictionary Language**: the default language of the dictionary (this is used in generating inflections of entries)
  - **Inflection lookup dictionary**: if LW supports the selected language, you can use the inflections dictionary to generate inflections. If not, you can specify your own dictionary.
  - **Default Part of Speech**: default part of speech of the entries you intend to add to the dictionary, you can change it for each entry at a later stage.
  - **Default Constraints**: (also referred to as BOFA) defines the compounding properties of the entries (composition could be using hyphens e.g. anti-social, apostrophes e.g. he's, or direct composition e.g. *head*ache).
  - **UIMA Type**: the *Type* (concept) that is used to name the occurrences of the dictionary entries when found in text. (Note how we added the prefix ***Dict*** to the type name, this is a good practice to distinguish dictionary types from others).
  - **UIMA Description**: a short description of the *Type.*



New Dictionary Database

Define Attributes of Dictionary Entries

Specify attributes of entries in this dictionary database.

Dictionary language: English

Inflection lookup dictionary: ☑ Use built-in dictionary

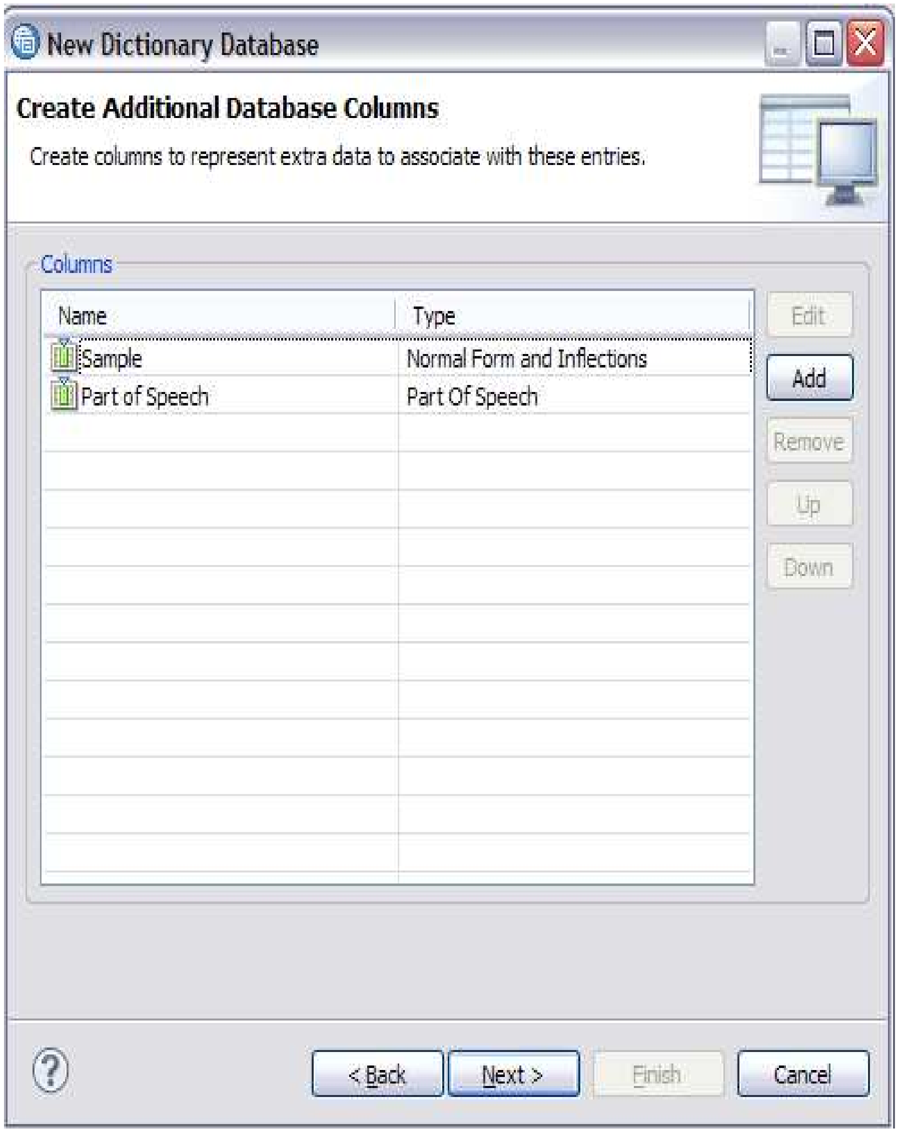
Default Part-of-Speech: Noun

Default Constraints: ☑ Alone ☑ Begin □ Mid ☑ End

UIMA type: com.ibm.DictFirstName

UIMA description: FirstName

- This dialog box shows you the column in the custom dictionary.
- The first column contains the name of the Type (here it is sample)
- The second column contains the part of speech (POS). The value will be the default you specified in the previous dialog box. It can be changed at a later stage.
- It is possible to add new columns called features (extra information linked to the dictionary entries). For example, if you have a dictionary with cities, you can add as many columns as you want (e.g. Population, longitude, latitude...). By clicking on the Add button, a dialog box pops up and you can configure the features (see next slide)

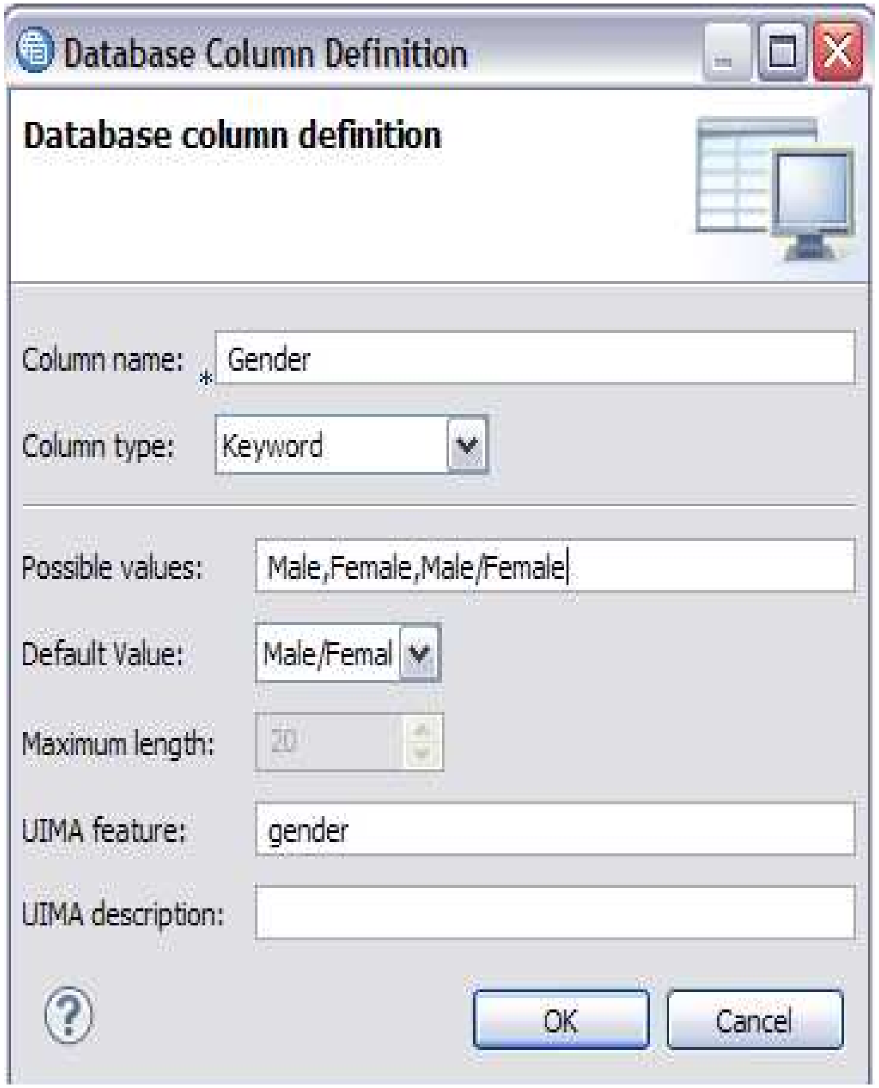


Create Additional Database Columns

Create columns to represent extra data to associate with these entries.

Name	Type
Sample	Normal Form and Inflections
Part of Speech	Part Of Speech

- Defining database columns (Features).
  - **Column name**: the name of the feature (for example, for a dictionary containing first names, you can add the feature gender).
  - **Column type**: The type of data to be stored as a value of the feature. It could be a string, a non-empty string, a boolean...). If the type of the feature is not respected, an error will be shown.
  - **Possible values**: a comma separated list of values of the feature, only applicable for Keyword type.
  - **Default value**: The default value to be used. It can be changed later when the entry is entered/edited.
  - **Maximum length**: of the feature value. It is better to be generous with the length to avoid the value to be truncated. The Maximum length can be increased.
  - **UIMA feature**: the name of the feature as it will appear in the output of the annotations.
  - **UIMA Description**: a short description of the feature (optional).


Database Column Definition

Database column definition

Column name: * Gender

Column type: Keyword

Possible values: Male,Female,Male/Female

Default Value: Male/Femal

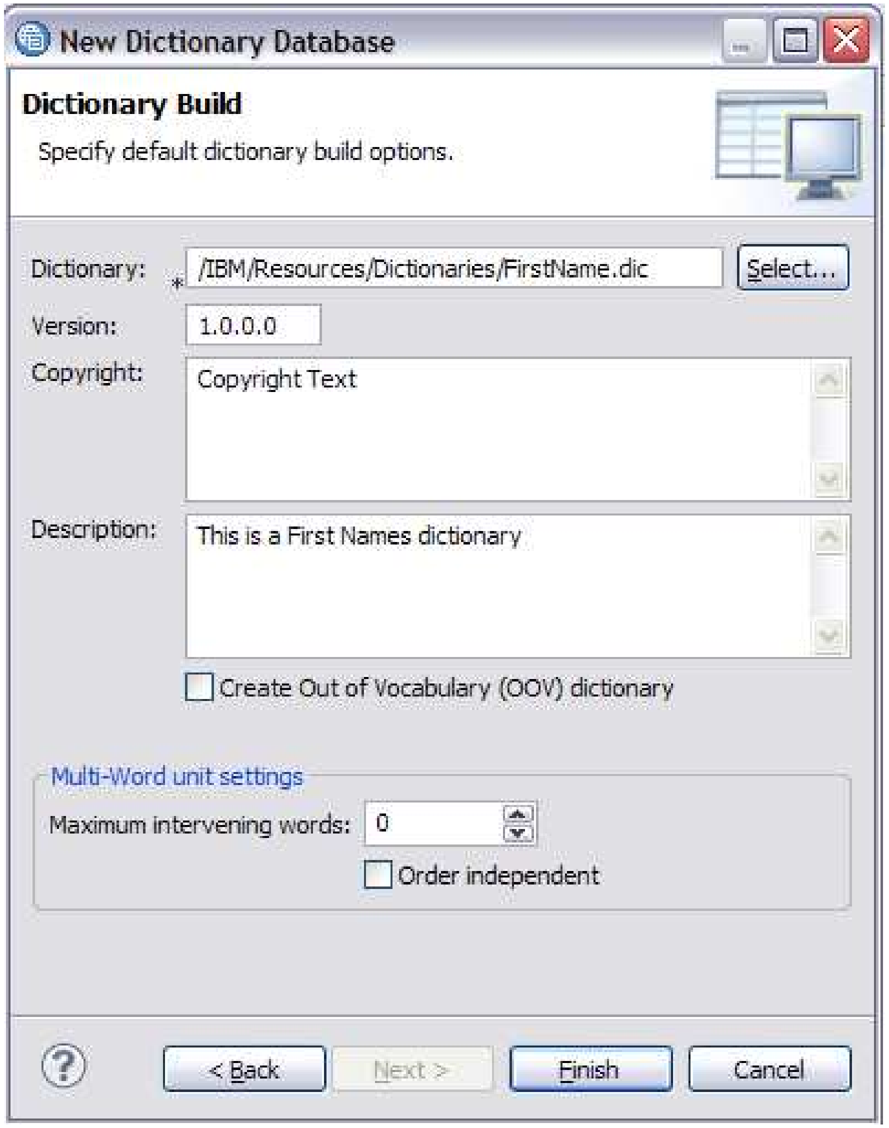
Maximum length: 20

UIMA feature: gender

UIMA description:

- Dictionary build configuration:
  - **Dictionary:** the dictionary name and build path.
  - **Version:** The dictionary version. The number will be increased every time the dictionary is built/compiled.
  - **Copyright:** Copyright statement filed.
  - **Description**: description of the dictionary
  - **Create OOV (Out Of Vocabulary) dictionary**: allows you to specify whether the dictionary created from the database will be an OOV dictionary. For more information on OOV dictionaries, please refer to the LW Help or Glossary)

- Multi-Word unit (MWU) settings:
  - **Maximum intervening words**: how many words will be accepted in between the words of a multi-word unit, for the term still to be matched. Default is “0”, which means it will only match if there are no intervening tokens between the MWU elements. If the value is 1 for instance, Alzheimer disease will also match Alzheimer's disease even though there is one intervening token, i.e., “**'s**”.
  - **Order independent**: The order of the MWU elements. If you check the box, then it will match even if the elements are not in the dictionary entry's order. Example, with the Maximum intervening words set to 1, and the order independent box checked, “Alzheimer's disease” and “disease of Alzheimer' will be both matched.



Dictionary Build

Specify default dictionary build options.

Dictionary: * /IBM/Resources/Dictionaries/FirstName.dic

Version: 1.0.0.0

Copyright: Copyright Text

Description: This is a First Names dictionary

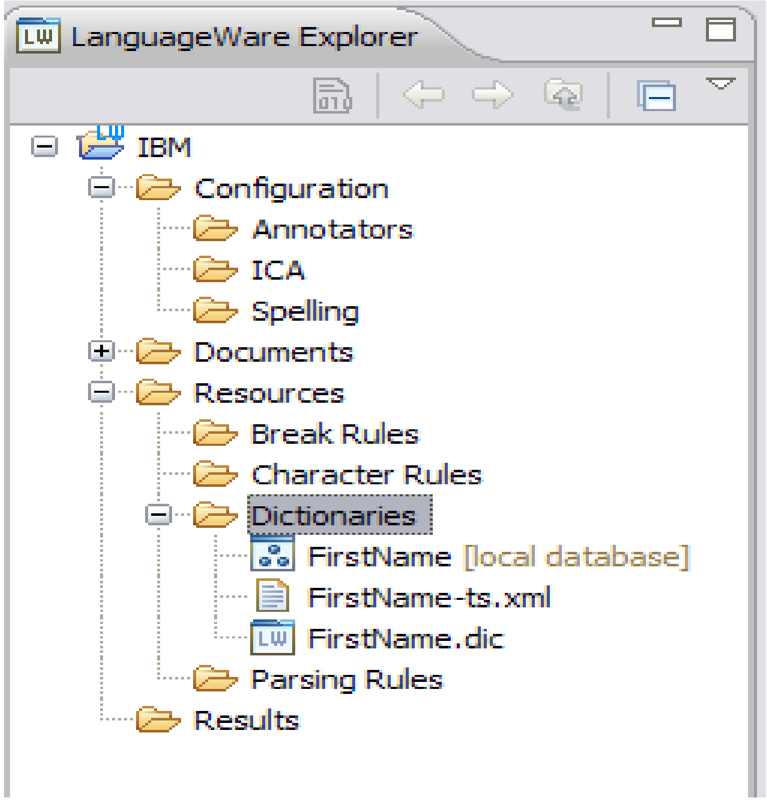
□ Create Out of Vocabulary (OOV) dictionary

Multi-Word unit settings

Maximum intervening words: 0

□ Order independent

- When you click finish, three elements are created:
  - **Database**: where the data is stored.
  - **Typesystem file (-ts.xml)**: an .xml file that stores the UIMA type(s) generated by the dictionary.
  - **Dictionary**: the compiled dictionary to be used in the annotator.



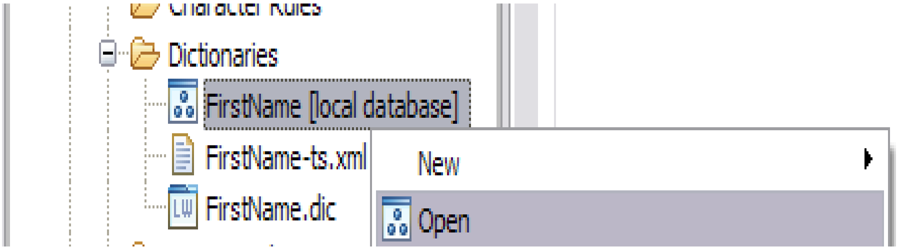
LanguageWare Explorer

- IBM
  - Configuration
    - Annotators
    - ICA
    - Spelling
  - Documents
  - Resources
    - Break Rules
    - Character Rules
    - Dictionaries
      - FirstName [local database]
      - FirstName-ts.xml
      - FirstName.dic
    - Parsing Rules
  - Results

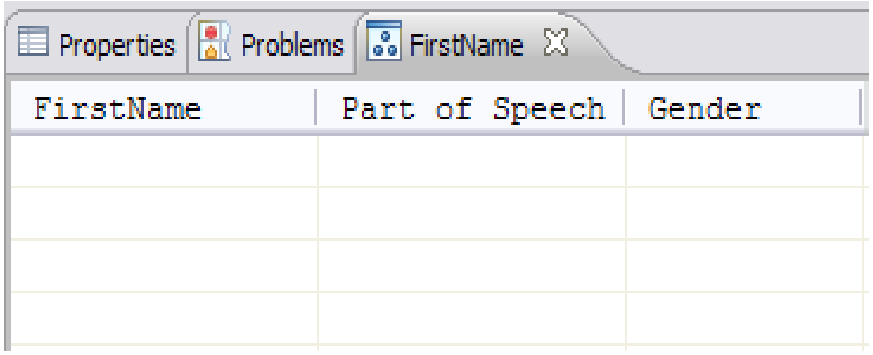
Adding entries to a custom dictionary

- To add an entry to a dictionary, it has to be added to the database, then the database is built/compiled.

- To add the entry, open the database:
  - Double click the database link on the LanguageWare explorer, or
  - Right click the database and select open.

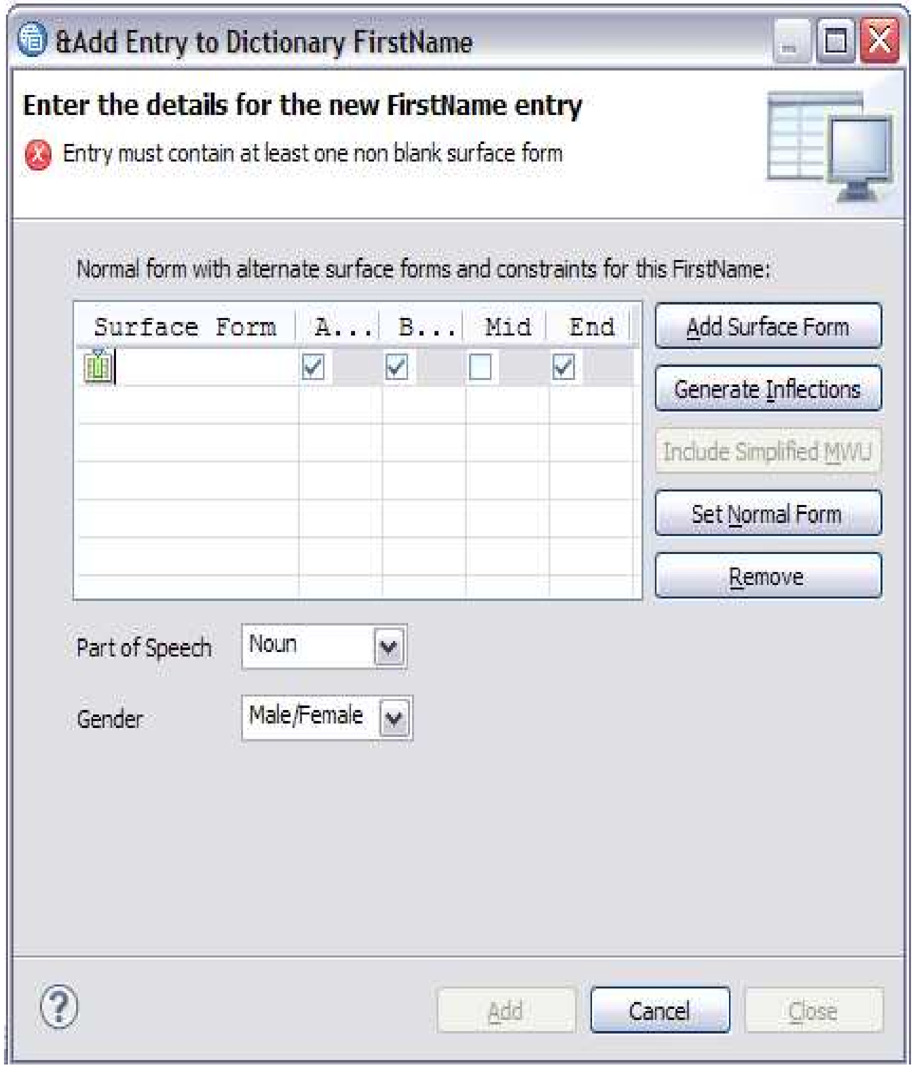


This will open the database in the “Database Editor”.



FirstName	Part of Speech	Gender

- Click the icon on the Database Editor tool bar, this will open the “add entry” window.
- **Add Surface Form**: surface form is the column that contains the different possible forms/inflections of the word. The icon indicates the normal form (lemma) of the entry.
- **Generate Inflections**: generates the possible inflections (when they are supported). For example, if you enter the noun “car”, and press click “Generate Inflections”, you will get “cars” added as other possible forms. You can also add other surface forms that would all map to the specified normal form.
- **Set Normal Form**: sets an entry as a normal form by selecting.
- **BOFA**: (Default constraints)values setting the compositional properties of a word. They are set by default when creating the dictionary database, but they can be changed.
- **Part of Speech**: of the entry.
- **Features**: the extra column(s) that were created in the database that store different information related to the entries; *Gender* in our example.
- **Tip**: you can also add entries to the database by selecting a word/phrase in the text open in the editor and dragging it into an open database in the database editor.



Enter the details for the new FirstName entry

Entry must contain at least one non blank surface form

Normal form with alternate surface forms and constraints for this FirstName:

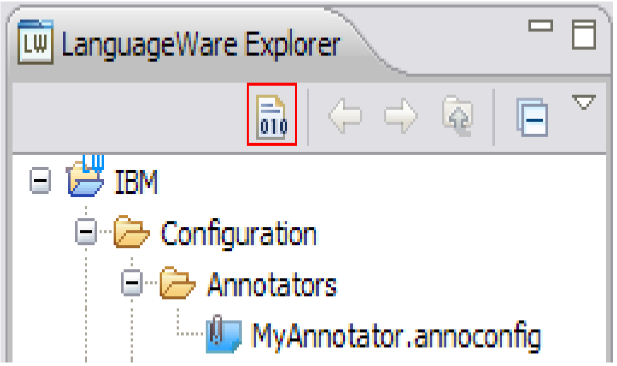
Surface Form	A...	B...	Mid	End
	☑	☑	□	☑

Part of Speech: Noun

Gender: Male/Female

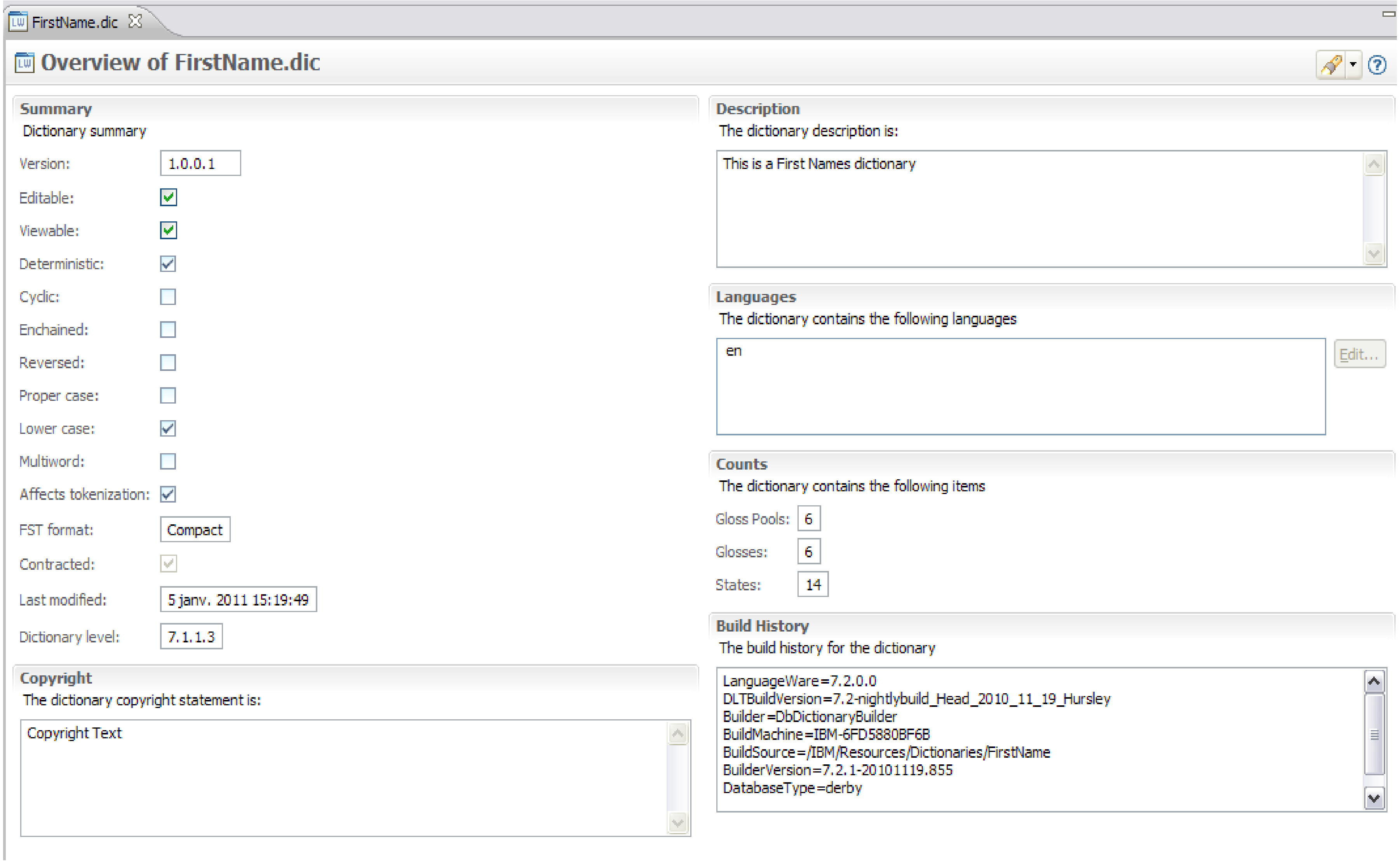
View a custom dictionary in the LanguageWare Editor

- When you finish adding entries to the dictionary, close the dialog box, then save the data and commit changes (by default, changes are committed automatically into the database).
- Build/compile the resource: select the database, right click and select build LanguageWare resources. Otherwise, you can simply select the database and click the 010 icon in the LanguageWare Explorer.



- Open the dictionary (double click on the ".dic" file) in the Editor view.

Dictionary overview in the LanguageWare Editor



Overview of FirstName.dic

Summary

Dictionary summary

Version: 1.0.0.1

Editable: ☑

Viewable: ☑

Deterministic: ☑

Cyclic: □

Enchained: □

Reversed: □

Proper case: □

Lower case: ☑

Multiword: □

Affects tokenization: ☑

FST format: Compact

Contracted: ☑

Last modified: 5 janv. 2011 15:19:49

Dictionary level: 7.1.1.3

Copyright

The dictionary copyright statement is:

Copyright Text

Description

The dictionary description is:

This is a First Names dictionary

Languages

The dictionary contains the following languages

en

Counts

The dictionary contains the following items

Gloss Pools: 6

Glosses: 6

States: 14

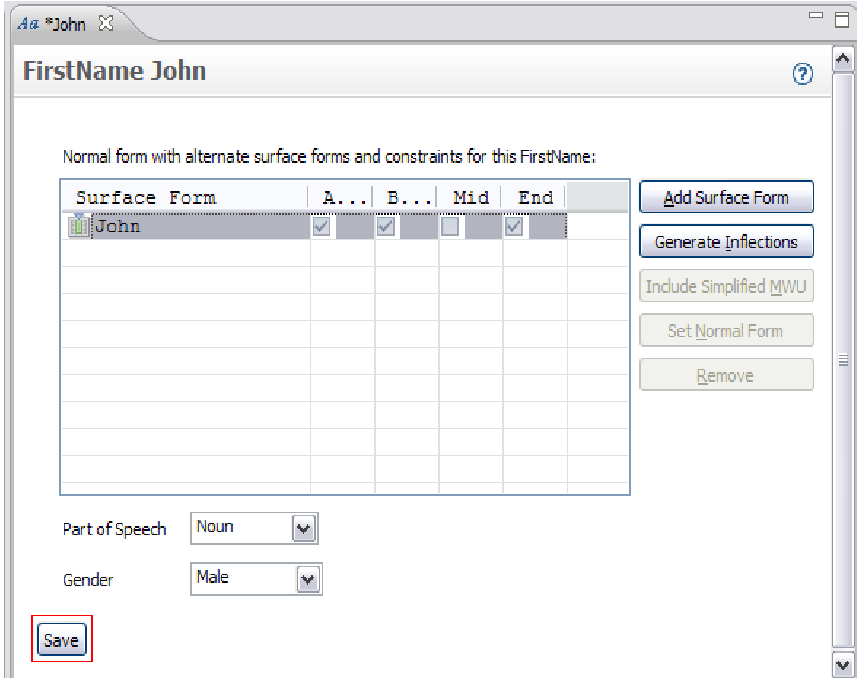
Build History

The build history for the dictionary

```
LanguageWare=7.2.0.0
DLTBuildVersion=7.2-nightlybuild_Head_2010_11_19_Hursley
Builder=DbDictionaryBuilder
BuildMachine=IBM-6FD5880BF6B
BuildSource=/IBM/Resources/Dictionaries/FirstName
BuilderVersion=7.2.1-20101119.855
DatabaseType=derby
```

Editing entries in the custom dictionary database.

- To edit entries in the custom dictionary, open the database, select the entry to be edited and double click on it. It will open in the LanguageWare Editor.
- Modify the entry attributes, then save it.
- Build/compile the dictionary for the changes to take effect.



FirstName John

Normal form with alternate surface forms and constraints for this FirstName:

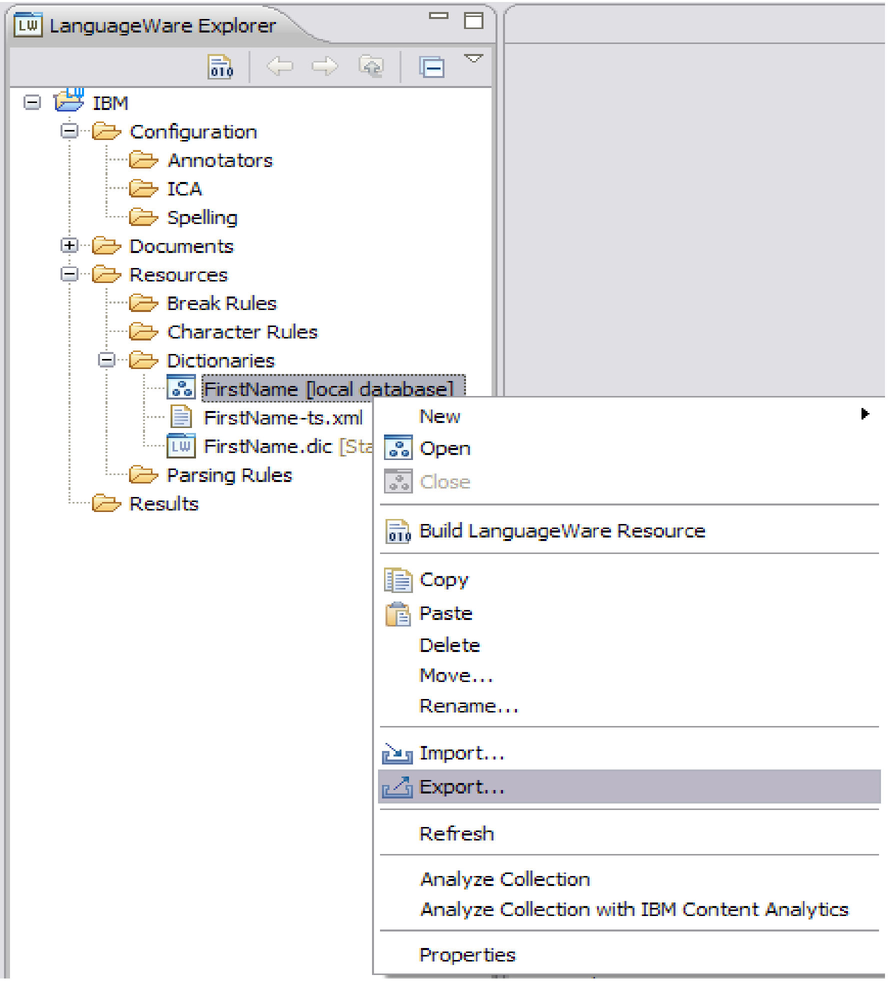
Surface Form	A...	B...	Mid	End
John	☑	☑	☐	☑

Part of Speech: Noun

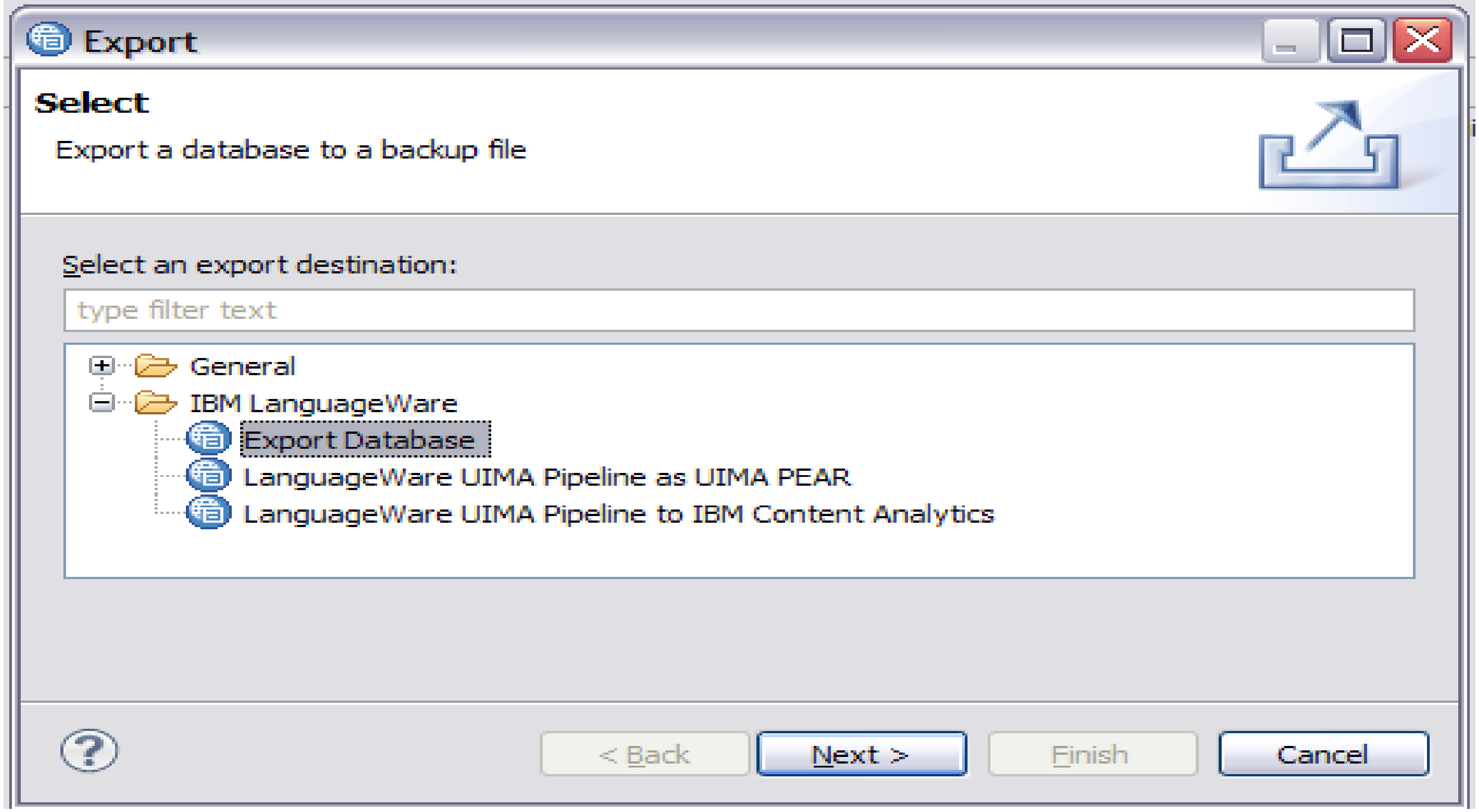
Gender: Male

Exporting a dictionary database

- Data is exported from the database in the form of a zip file, containing ***.csv*** files. The file containing the data is the DATA.csv.
- Make sure the database is closed (right click, select close),
- Right-click the database link in the LanguageWare Explorer and select export: a dialog box will pop up.



- Select the “IBM LanguageWare/Export Database” option, then click Next



Export

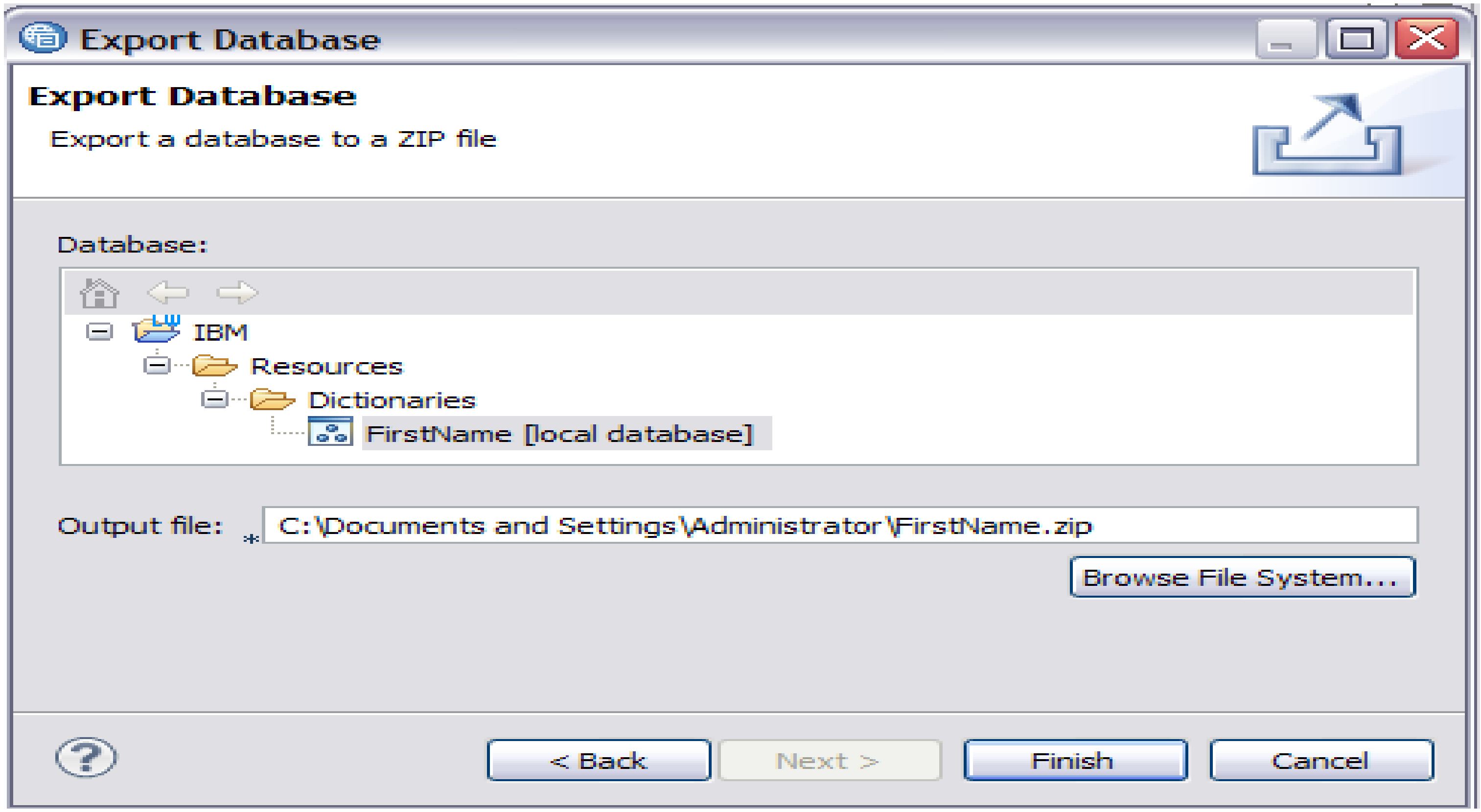
Select

Export a database to a backup file

Select an export destination:

type filter text
General
IBM LanguageWare
Export Database
LanguageWare UIMA Pipeline as UIMA PEAR
LanguageWare UIMA Pipeline to IBM Content Analytics
< Back
Next >
Finish
Cancel

- Select the Database to be exported, select the output file path and name, then click Finish.



Export Database

Export Database

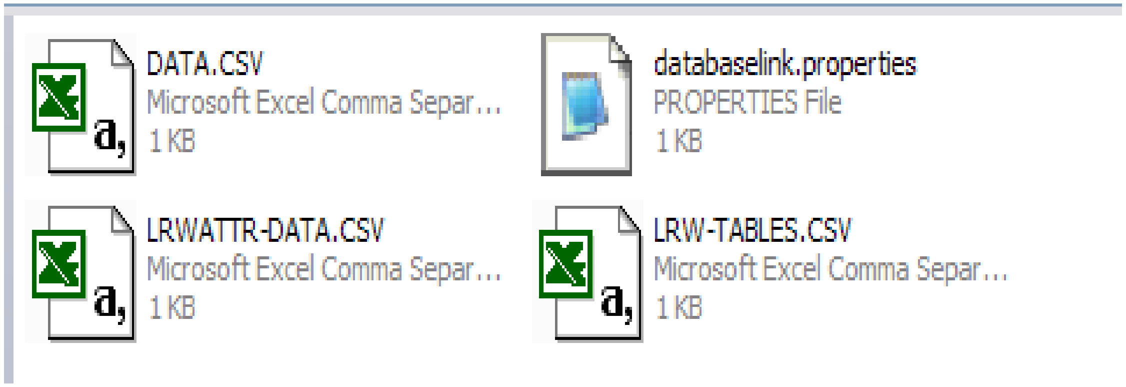
Export a database to a ZIP file

Database:

IBM
Resources
Dictionaries
FirstName [local database]

Output file: C:\Documents and Settings\Administrator\FirstName.zip

- If you open the exported .zip file, you can see the different .csv files. The DATA.CSV file contains the dictionary data. The remaining files are database tables and properties.



DATA.CSV
Microsoft Excel Comma Separ...
1 KB

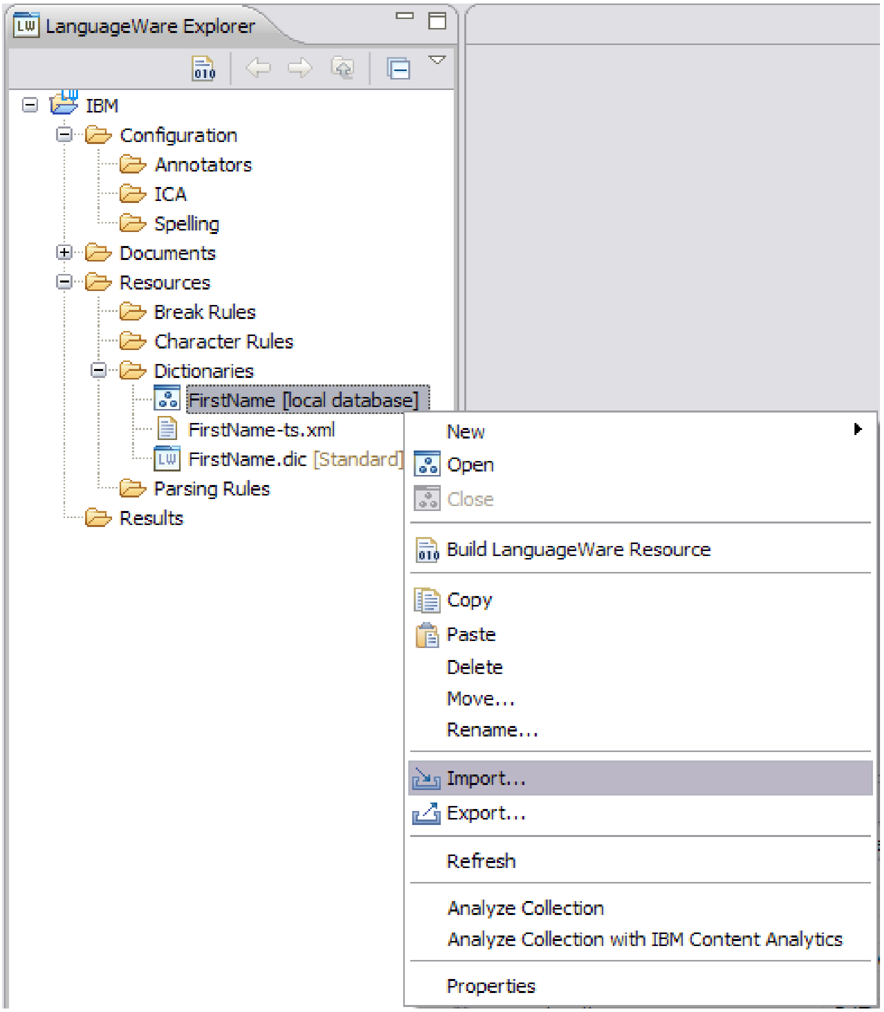
databaselink.properties
PROPERTIES File
1 KB

LRWATTR-DATA.CSV
Microsoft Excel Comma Separ...
1 KB

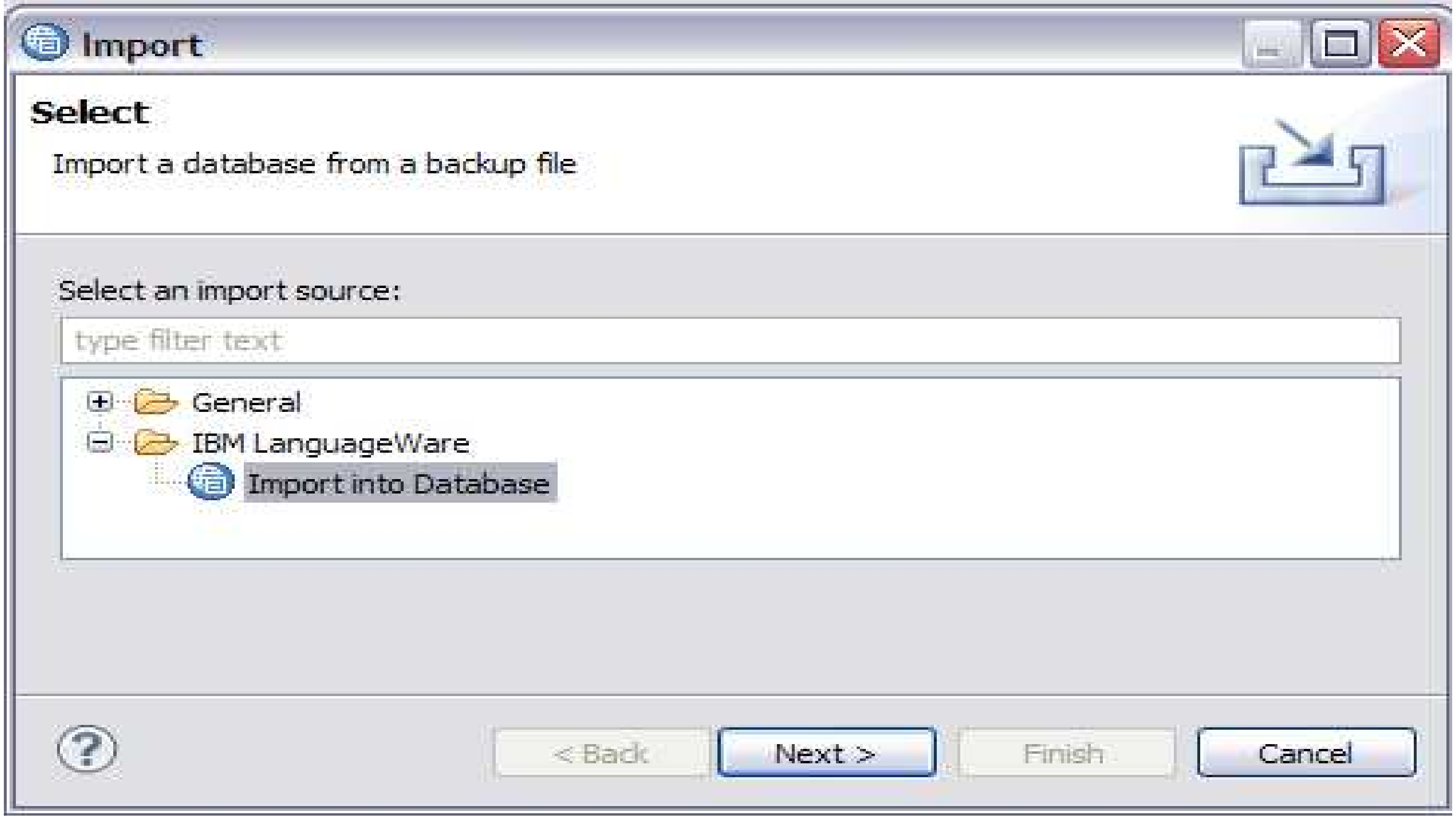
LRW-TABLES.CSV
Microsoft Excel Comma Separ...
1 KB

Importing data into a dictionary database

- Data can be imported into a database in the form of a ***.csv*** files (comma separated document).
- Make sure the database is closed (right click, select close),
- Right-click the database link in the LanguageWare Explorer and select import: a dialog box will pop up.



- Select the “IBM LanguageWare/Import into Database” option, then click Next



Import

Select

Import a database from a backup file

Select an import source:

type filter text

- Input File: must be a ***.csv*** file.

- **File Encoding**: the default is UTF-8.

- **Surface form separator**: in the **.csv** file, the surface forms are in the same column as the lemma, but they are separated with a character different from the comma.

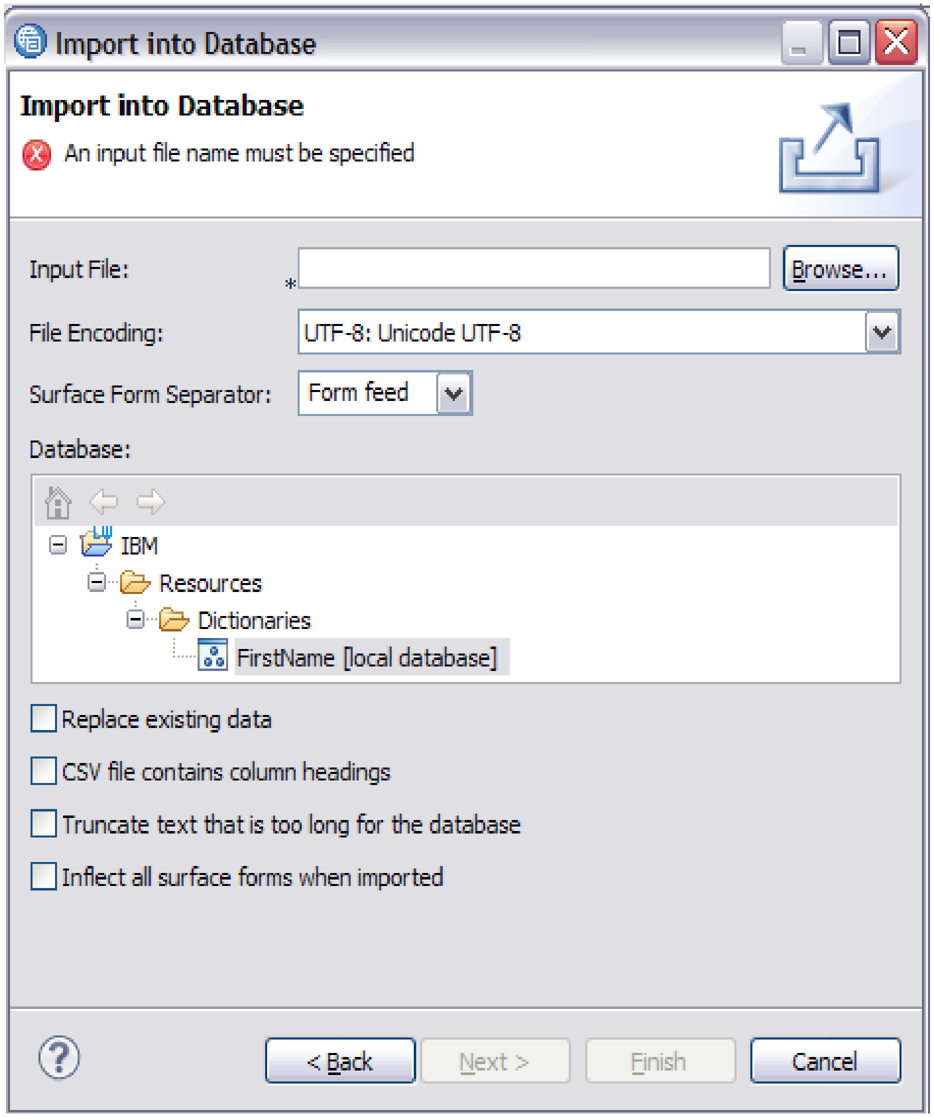
- **Database**: select the database you want to import the data into.

- **Replace existing data**: check this box if you want to overwrite the data in the database with the new imported data.

- **CSV contains column headings**: check this box if the ***.csv*** files contains headings (especially if it is an export from another custom database).

- **Truncate text that is too long for the database**: if you check this box, the entries that are too long for the default column size will be truncated.

- **Inflect surface forms when imported**: check this box to inflect the entries of the imported file. Inflections will be based on part of speech, language support for generating inflected forms, etc..



Import into Database

Import into Database

An input file name must be specified

Input File: *

File Encoding: UTF-8: Unicode UTF-8

Surface Form Separator: Form feed

Database:

- IBM
  - Resources
    - Dictionaries
      - FirstName [local database]

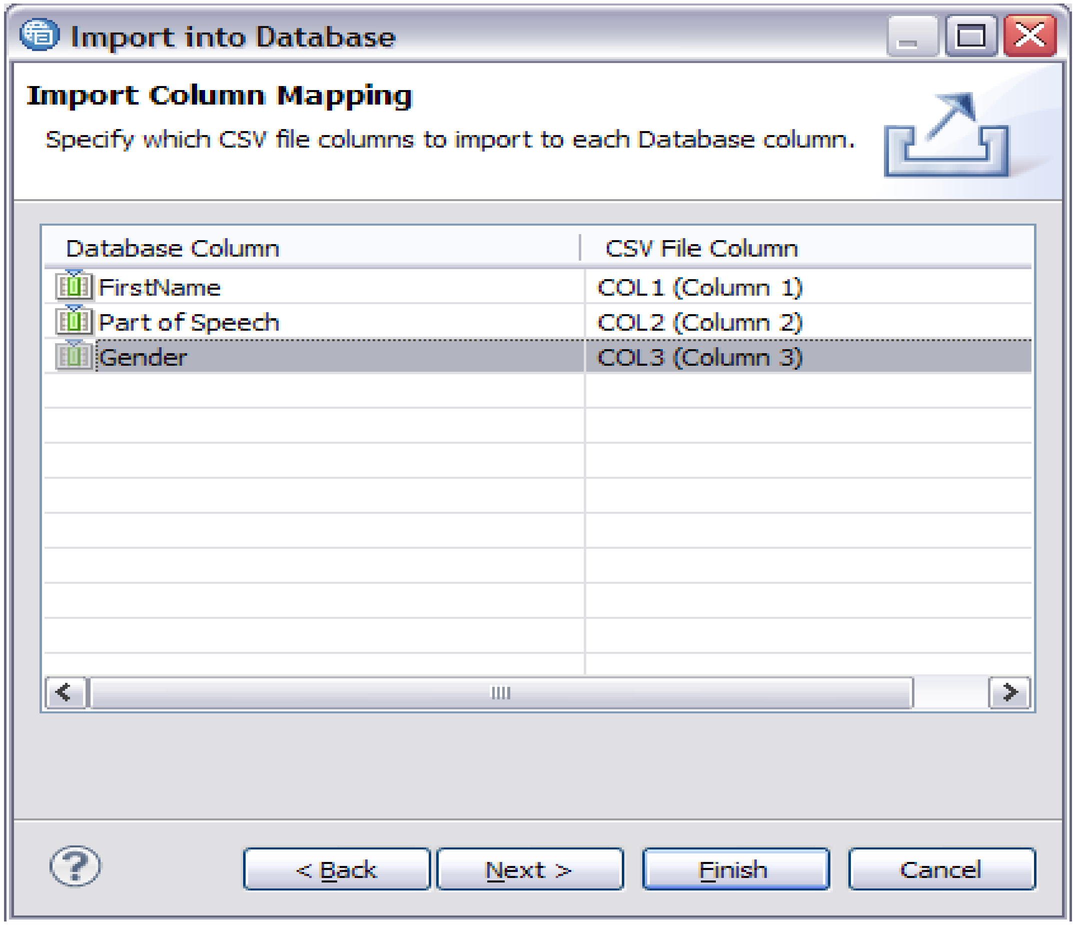
☐ Replace existing data

☐ CSV file contains column headings

☐ Truncate text that is too long for the database

☐ Inflect all surface forms when imported

- This dialog box maps the columns of the export file to the columns in the database. You can change the mapping by changing the values in the “CSV File Column”. Click on the value of the column and choose from the menu.



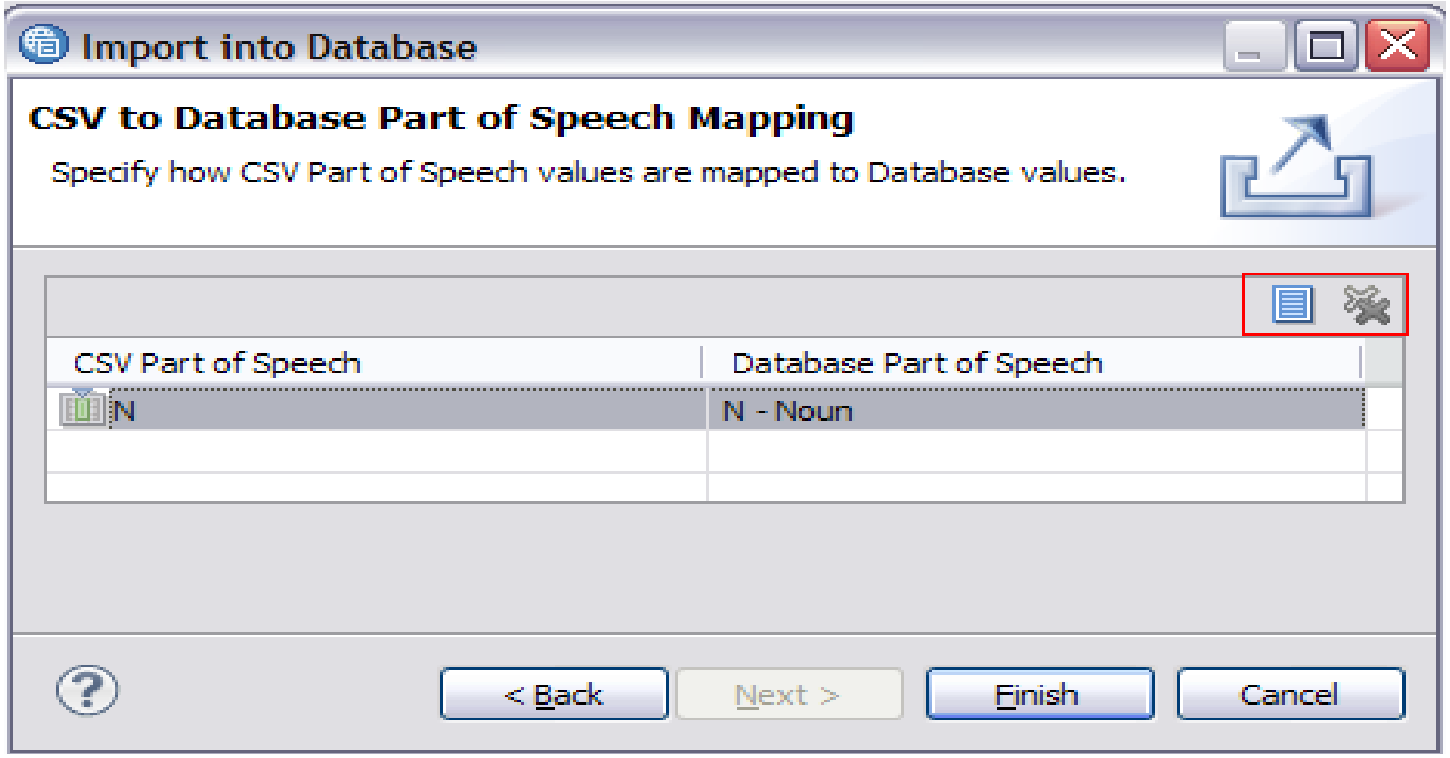
Import into Database

Import Column Mapping

Specify which CSV file columns to import to each Database column.

Database Column	CSV File Column
FirstName	COL1 (Column 1)
Part of Speech	COL2 (Column 2)
Gender	COL3 (Column 3)

- This dialog box maps the part of speech in the import file to those of the database. You can delete or modify the mapping of the CSV file using the icons on the right hand side.
- Click Finish to complete the import and then build/compile the database.



Import into Database

CSV to Database Part of Speech Mapping

Specify how CSV Part of Speech values are mapped to Database values.

CSV Part of Speech	Database Part of Speech
N	N - Noun

Module roadmap

- Custom dictionary database
  - What is it?
  - How to configure it?
  - How to add, edit, export and import data?
- **Summary and best practices**
- Sample exercises

Module summary

You have completed this module and can:

- Create and configure a custom dictionary database
- Add and modify entries to a custom dictionary database
- export and import data from and into a custom dictionary

See the LanguageWare help for more tips and advanced use cases.

Best practices

- A custom dictionary is a placeholder of terms belonging to a specific concept (Type). It could be car brands, product names, countries, first names ...
- A dictionary entry has a normal form (a lemma to which all the inflections/possible forms will be mapped), a BOFA (compositional properties of the form), and features (relevant information added as needed).
- When naming a dictionary, make sure the name explicitly represents the concept.
- When naming a dictionary UIMA Type, it is good practice to add the prefix “Dict” to the Type name as it will distinguish the types generated from dictionaries from those generated from rules.
- Use “Features' to get a better classification/abstraction of the dictionary entries, and also to have more flexibility when creating rules.
- Case is important when using the dictionary in a UIMA annotator:
  - if the entry is lower case, it will match lower case, Title Case and UPPER CASE.
  - if the entry is Title Case, it will match Title Case and UPPER CASE.
  - if the entry is UPPER CASE, it will only match UPPER CASE.

Module roadmap

- Custom dictionary database
  - What is it?
  - How to configure it?
  - How to add, edit, export and import data?
- Summary and best practices
- **Sample exercises**

Practice exercises

- Create a dictionary of Flavors
- Add a few words to the dictionary (chocolate, vanilla, mint, orange).
- Import a CSV file containing more flavors (an exported Database, OtherFlaovors is provided in Annex.zip so you can use it to add more data to your existing database).

Contacts

- If you have any questions, comments or suggestions, contact us using the LanguageWare email address *EMEALAN@ie.ibm.com* or on the developerWorks® Forum.

Trademarks, copyrights, and disclaimers

THE INFORMATION CONTAINED IN THIS PRESENTATION IS PROVIDED FOR INFORMATIONAL PURPOSES ONLY. WHILE EFFORTS WERE MADE TO VERIFY THE COMPLETENESS AND ACCURACY OF THE INFORMATION CONTAINED IN THIS PRESENTATION, IT IS PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED. IN ADDITION, THIS INFORMATION IS BASED ON IBM’S CURRENT PRODUCT PLANS AND STRATEGY, WHICH ARE SUBJECT TO CHANGE BY IBM WITHOUT NOTICE. IBM SHALL NOT BE RESPONSIBLE FOR ANY DAMAGES ARISING OUT OF THE USE OF, OR OTHERWISE RELATED TO, THIS PRESENTATION OR ANY OTHER DOCUMENTATION. NOTHING CONTAINED IN THIS PRESENTATION IS INTENDED TO, NOR SHALL HAVE THE EFFECT OF, CREATING ANY WARRANTIES OR REPRESENTATIONS FROM IBM (OR ITS SUPPLIERS OR LICENSORS), OR ALTERING THE TERMS AND CONDITIONS OF ANY AGREEMENT OR LICENSE GOVERNING THE USE OF IBM PRODUCTS OR SOFTWARE.