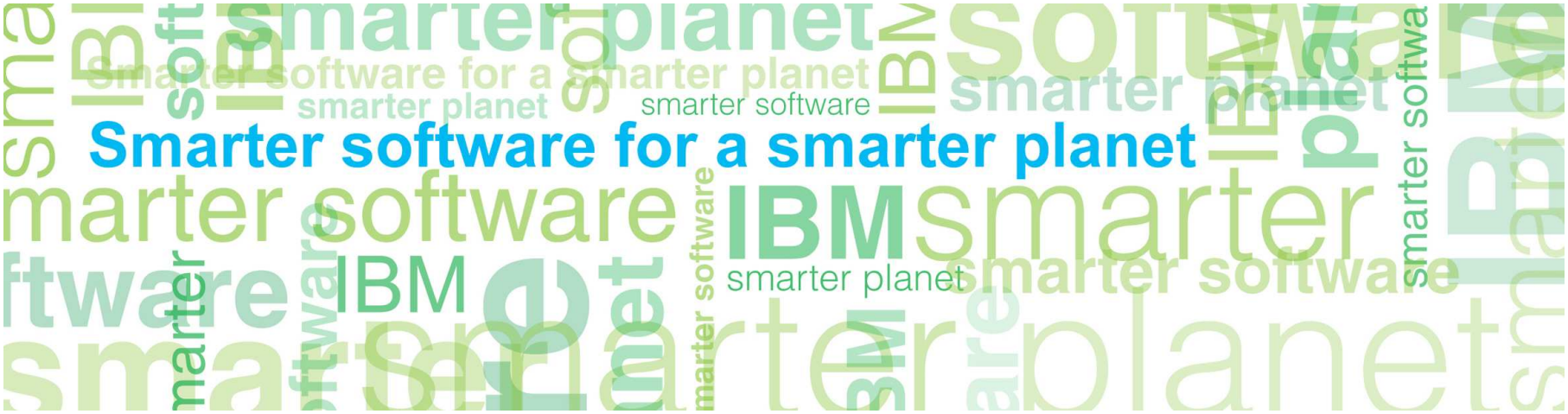
LanguageWare Resource Workbench 7.2

Create a UIMA pipeline configuration



Introduction

- **Module overview**
  - How to create and configure a UIMA pipeline.
  - Best practices
- **Target audience:**
  - All audiences
- **Prerequisites:**
  - Install LRW
  - Create a project,
  - Create dictionaries and parsing rules databases
- **Version release date** LRW 7.2, ICA 2.2, released october, 2010

Module objectives

After this module you will be able to:

- Create a UIMA pipeline configuration
- Run a UIMA Pipeline configuration on documents and see the results in the outline

Module roadmap

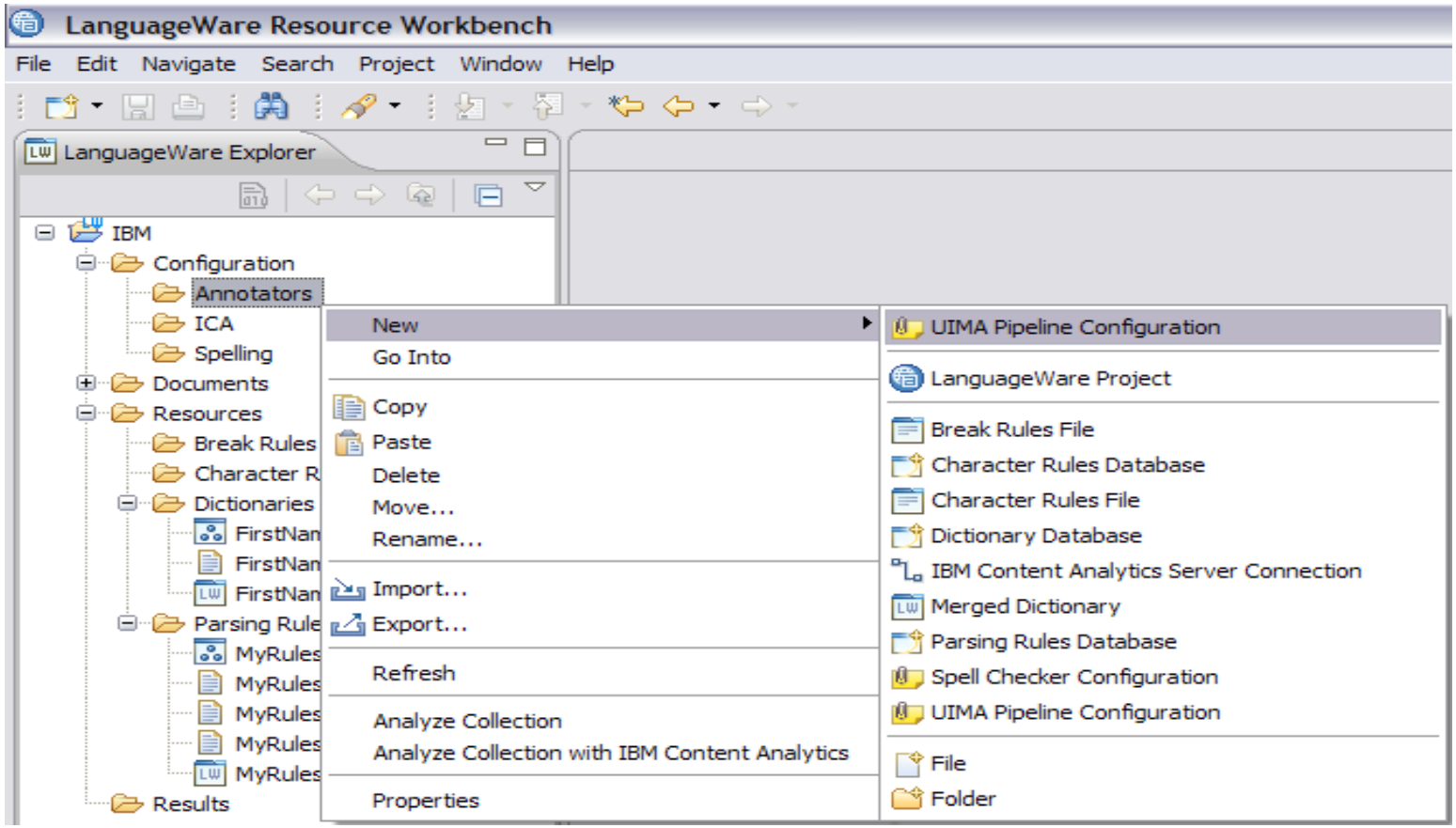
- **UIMA pipeline configuration**
  - What is it?
  - How to configure it and run it?
  - How to see the annotation results?
- **Summary and best practices**
- **Sample exercises**

UIMA pipeline configuration
What is it?

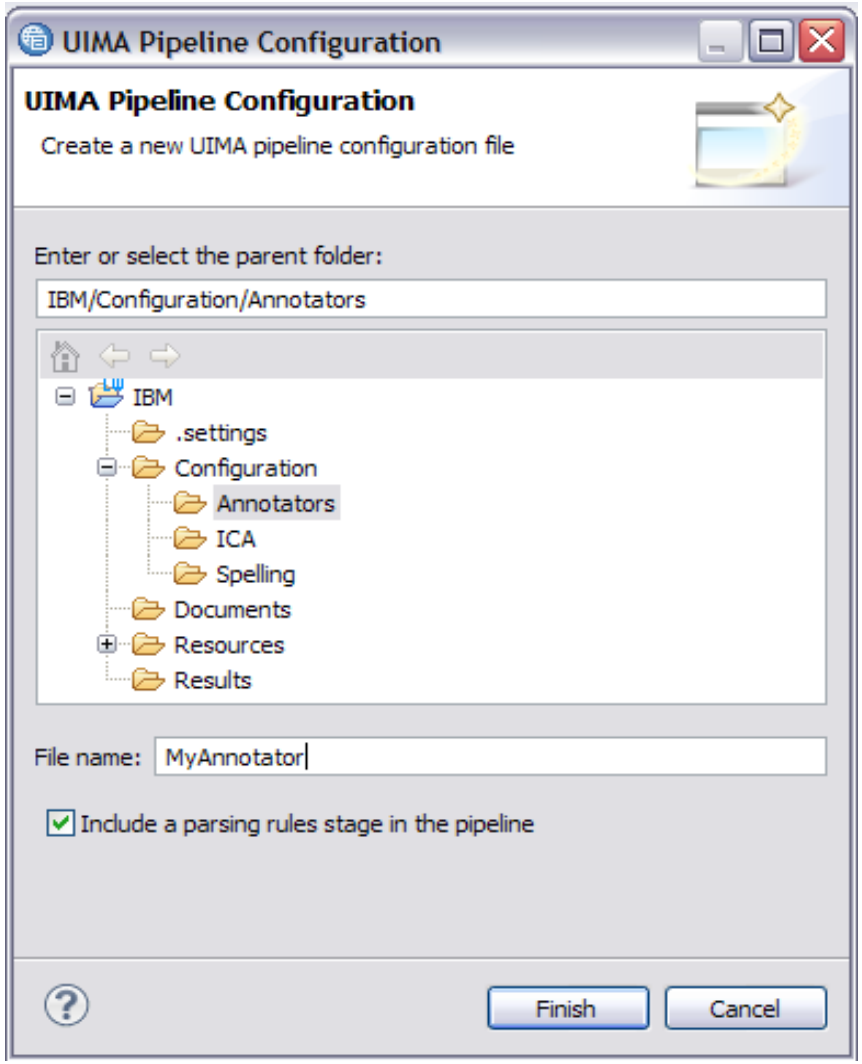
- **General**
  - *The UIMA pipeline configuration file is a place holder for the resources used to annotate documents.*
  - *It contains different stages (language, lexical analysis, parsing rules and clean up).*
  - *The stages are run consecutively and interact together to generate annotations.*
  - *We will refer to the UIMA pipeline configuration as UIMA pipeline in this tutorial.*

UIMA pipeline configuration
how to create and configure it?

- *In the LanguageWare® Explorer, right-click on Project/Configuration\Annotators, select New/ UIMA Pipeline Configuration*

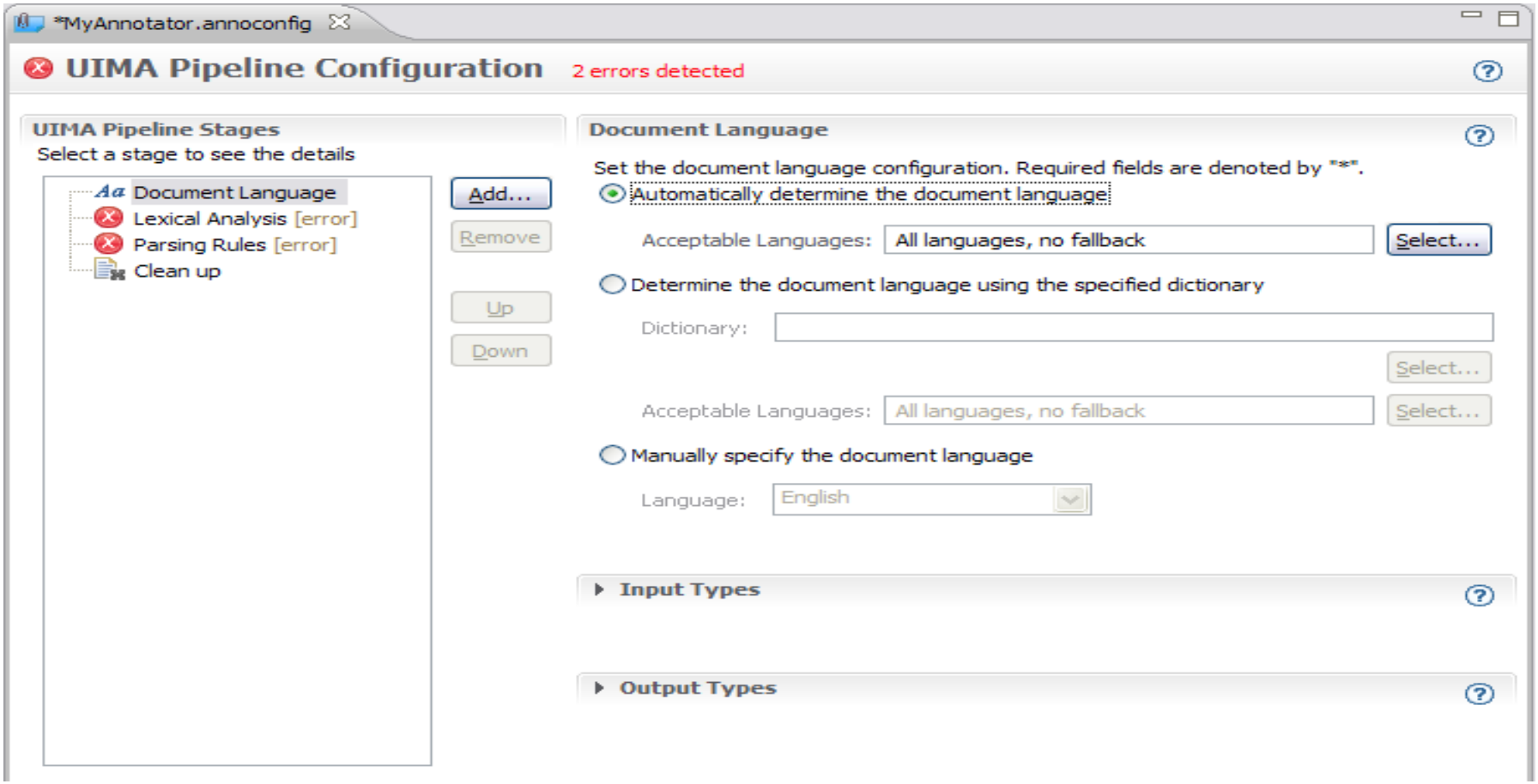


- Name the UIMA Pipeline configuration. It's good practice to use comprehensive names (related to the model or the general purpose of the annotator, so it could be something like Find Addresses, Police Report Analyzer, etc...)
- The “Include parsing rules stage in the pipeline” option is checked by default. It will create a parsing Rules stage by default in the pipeline. If you need an annotator that only uses dictionaries, uncheck this box. You can add a parsing rules stage later.
- Click finish, this will create an **.annoconfig** file.



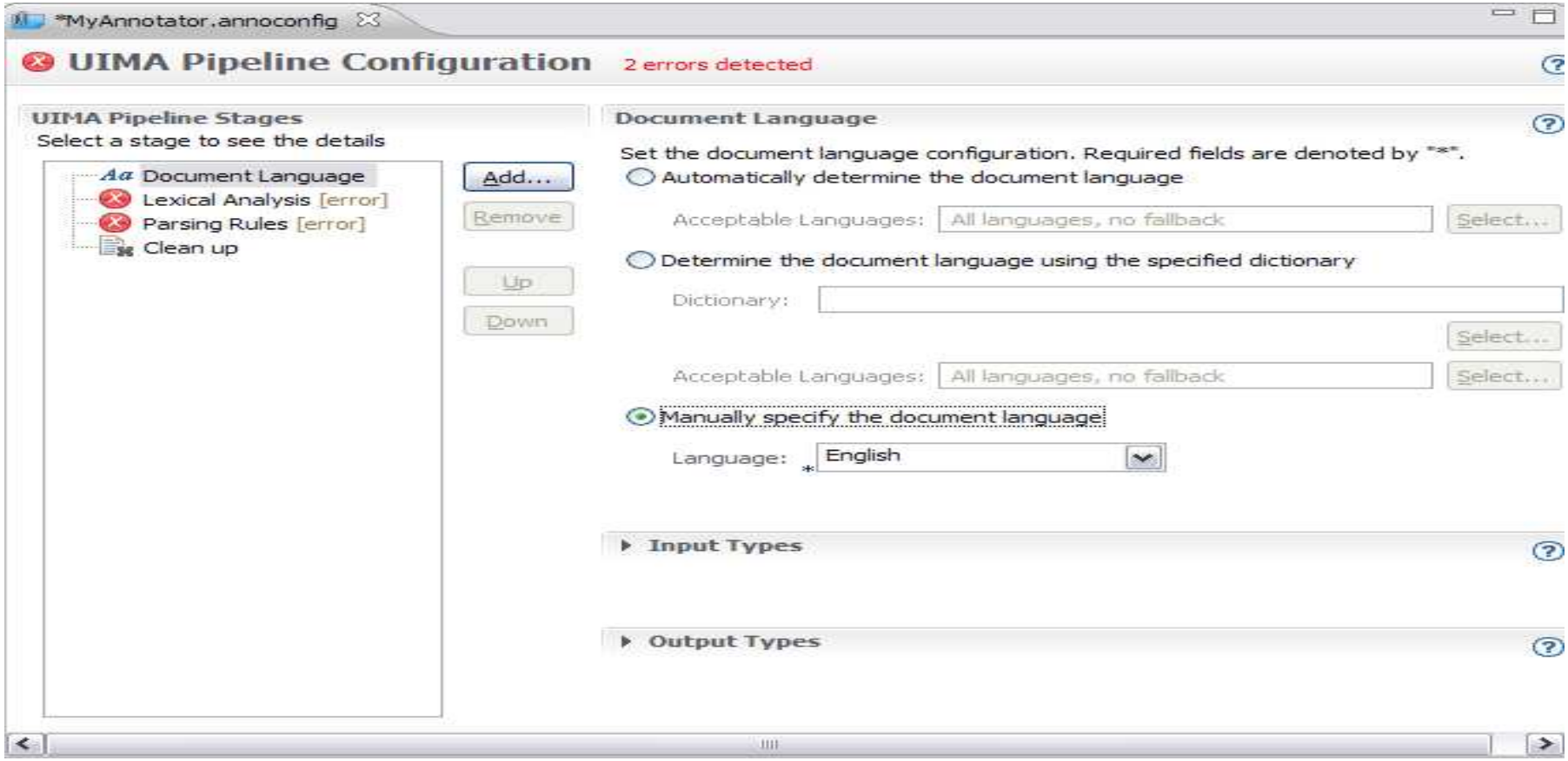
The UIMA pipeline configuration stages

- The UIMA Pipeline configuration has 4 main stages (Document Language, Lexical Analysis, Parsing Rules and Clean up stage). The pipeline will not run unless the proper resources have been added, that's why you have the error message.



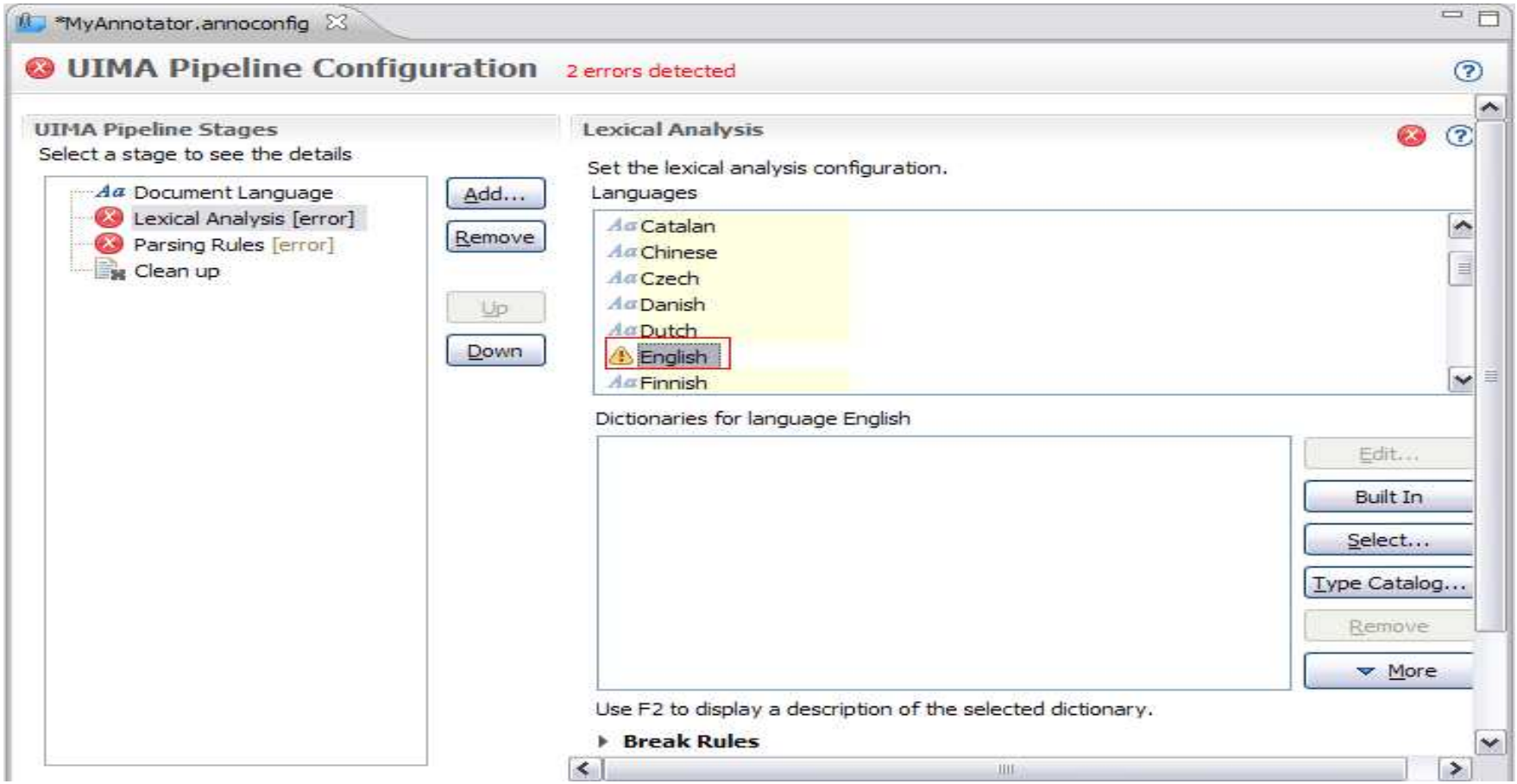
Document language stage

- For the aim of the training we will use the “Manually Specify the document language” option and we will set it to English. For advanced options, please refer to the LanguageWare Help.

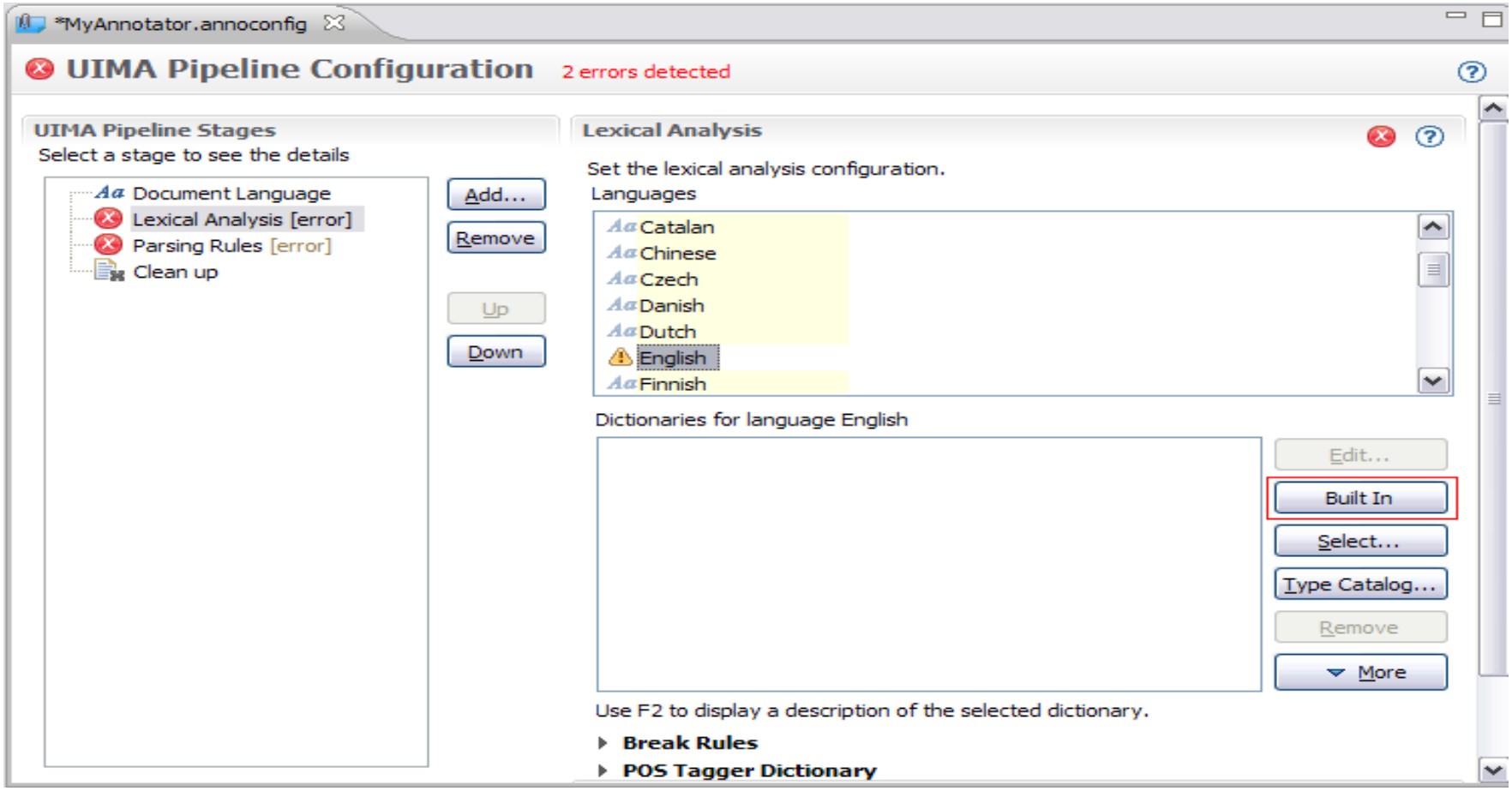


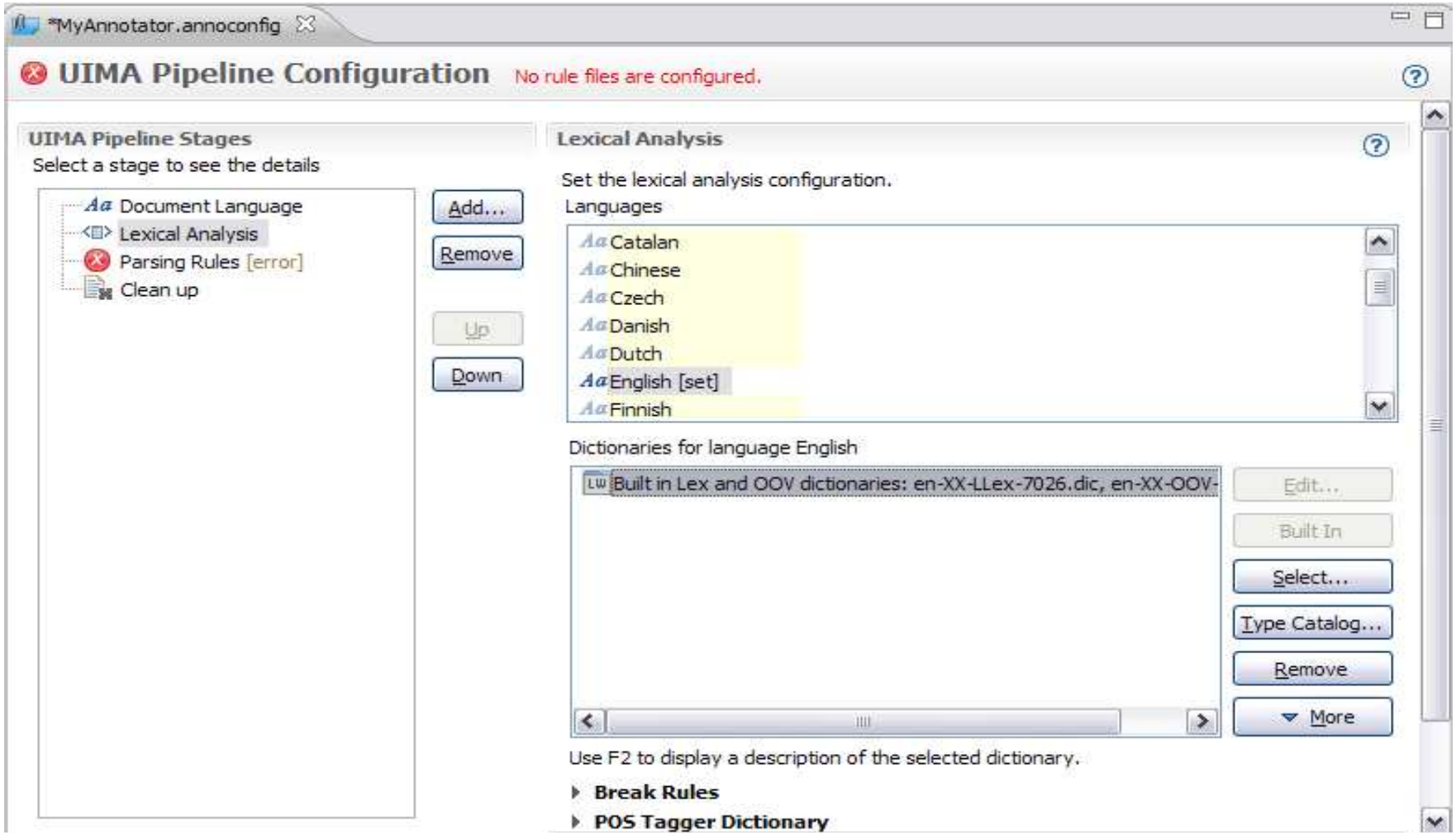
Lexical analysis stage

- In the Lexical Analysis stage, you specify the dictionary resources that will be used in the pipeline. Make sure the language is set properly. If you selected English in the “Document Language” stage, it will be set accordingly in the Languages tab.

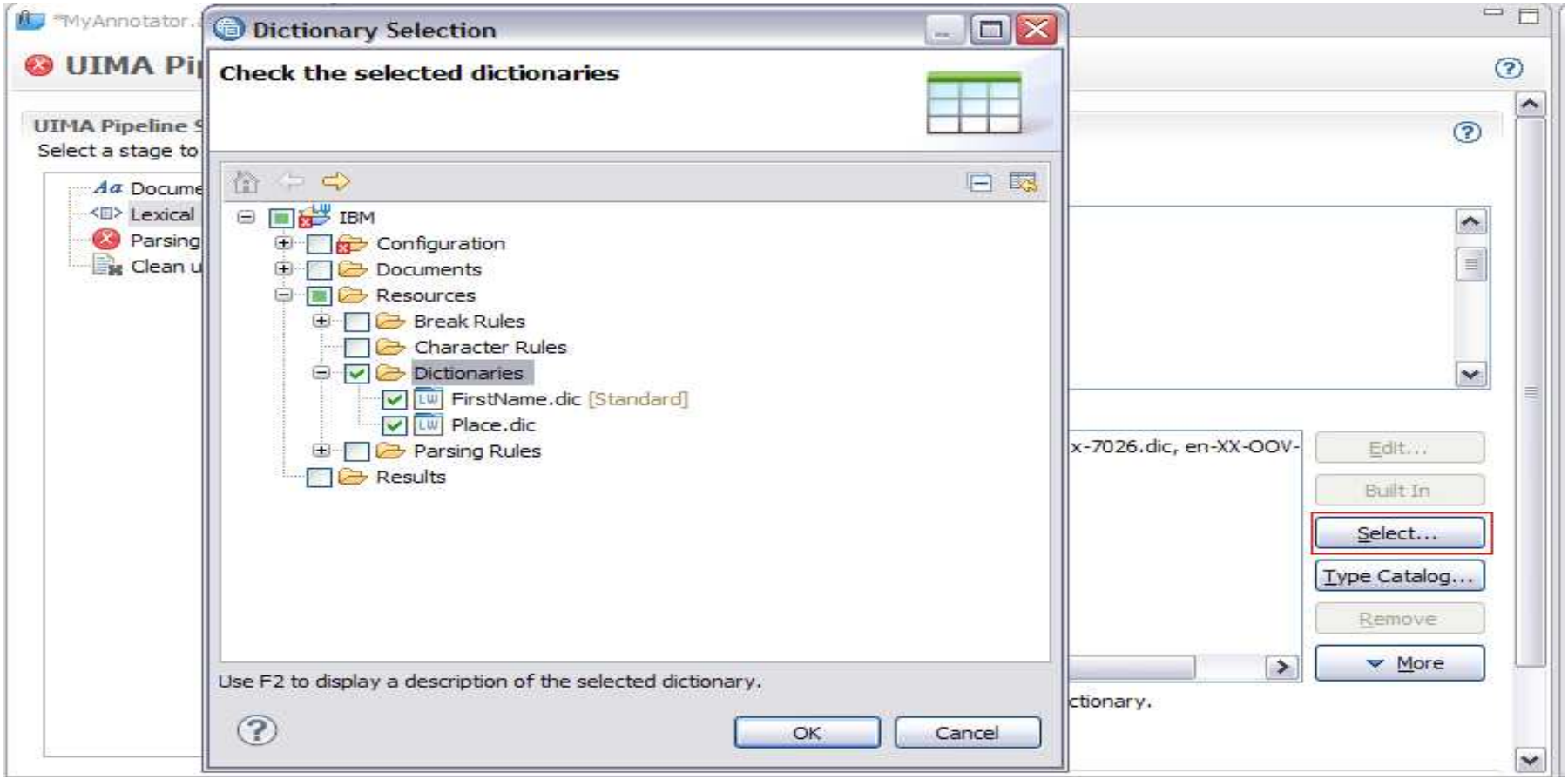


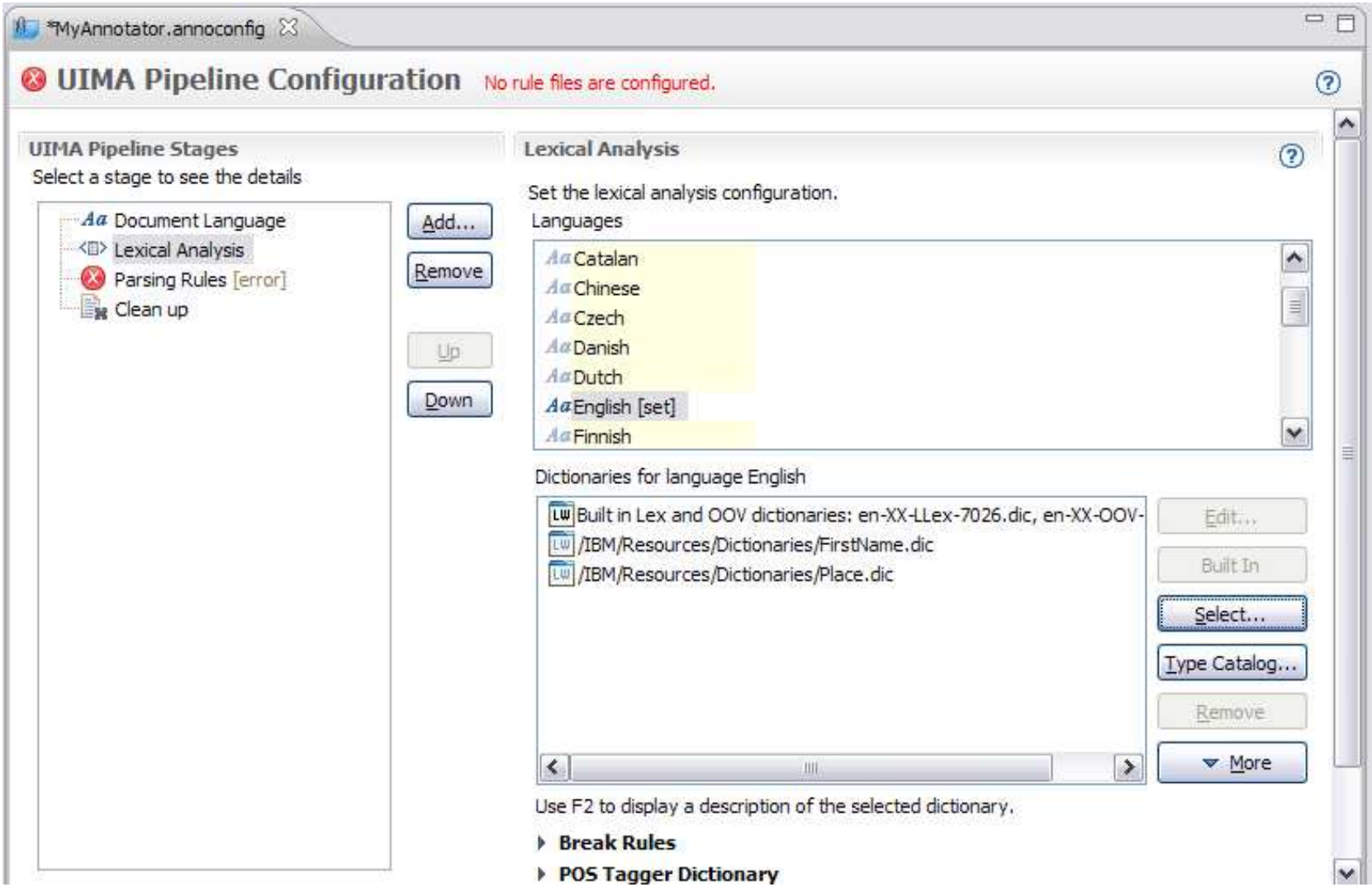
- Built in dictionaries for a number of languages are provided with LanguageWare and they cover Lexical Analysis (Lex) and Out-Of-Vocabulary (OOV).
- If you have a different Lex dictionary you can use it by ignoring the “Built In” and adding it using the Select button.





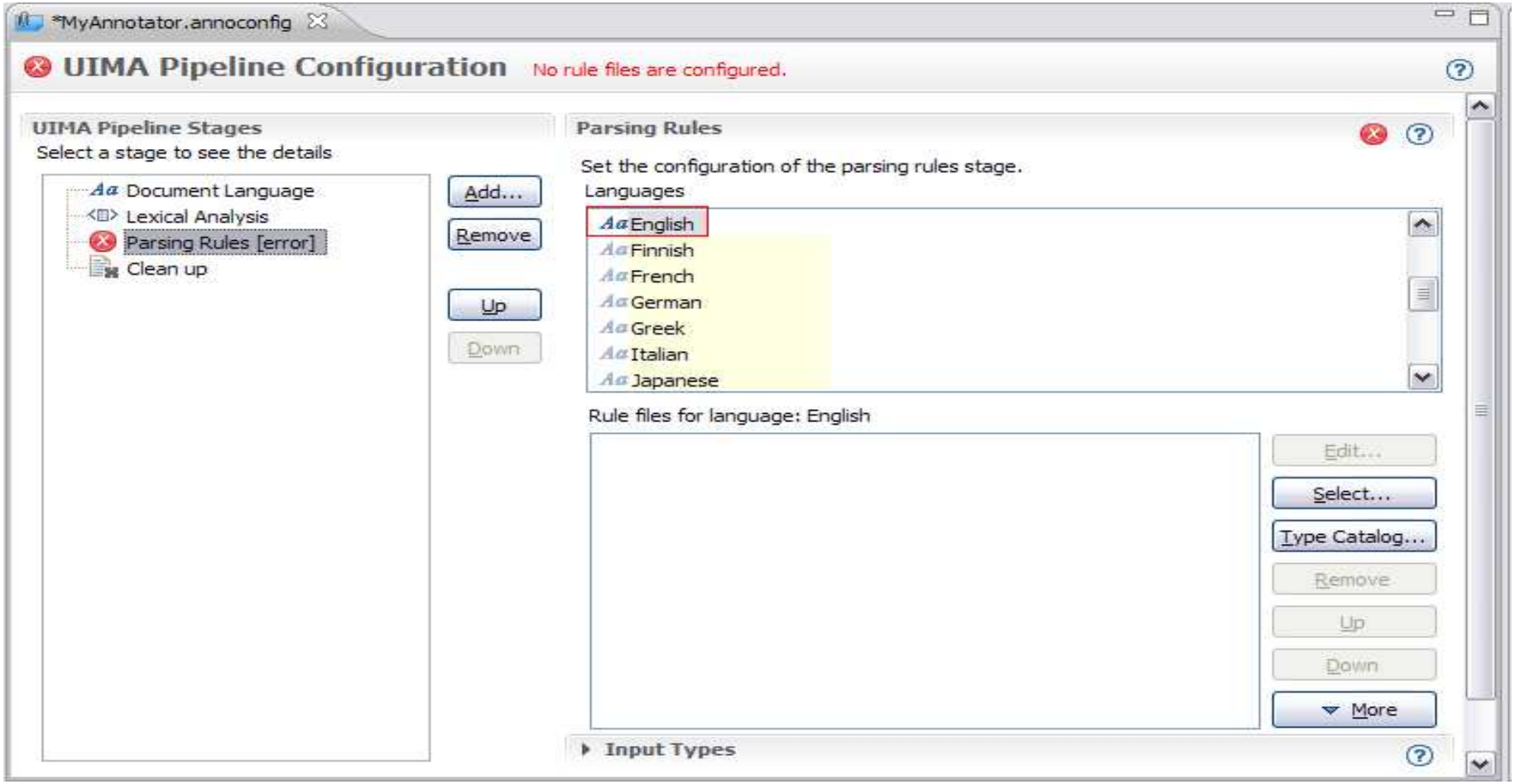
- To add custom dictionaries, click the Select button, this will open the workspace browser allowing you to add the relevant resources to the pipeline. Click OK after you made your selection.
- You can also drag the dictionaries from the LanguageWare Explorer into the dictionaries field.



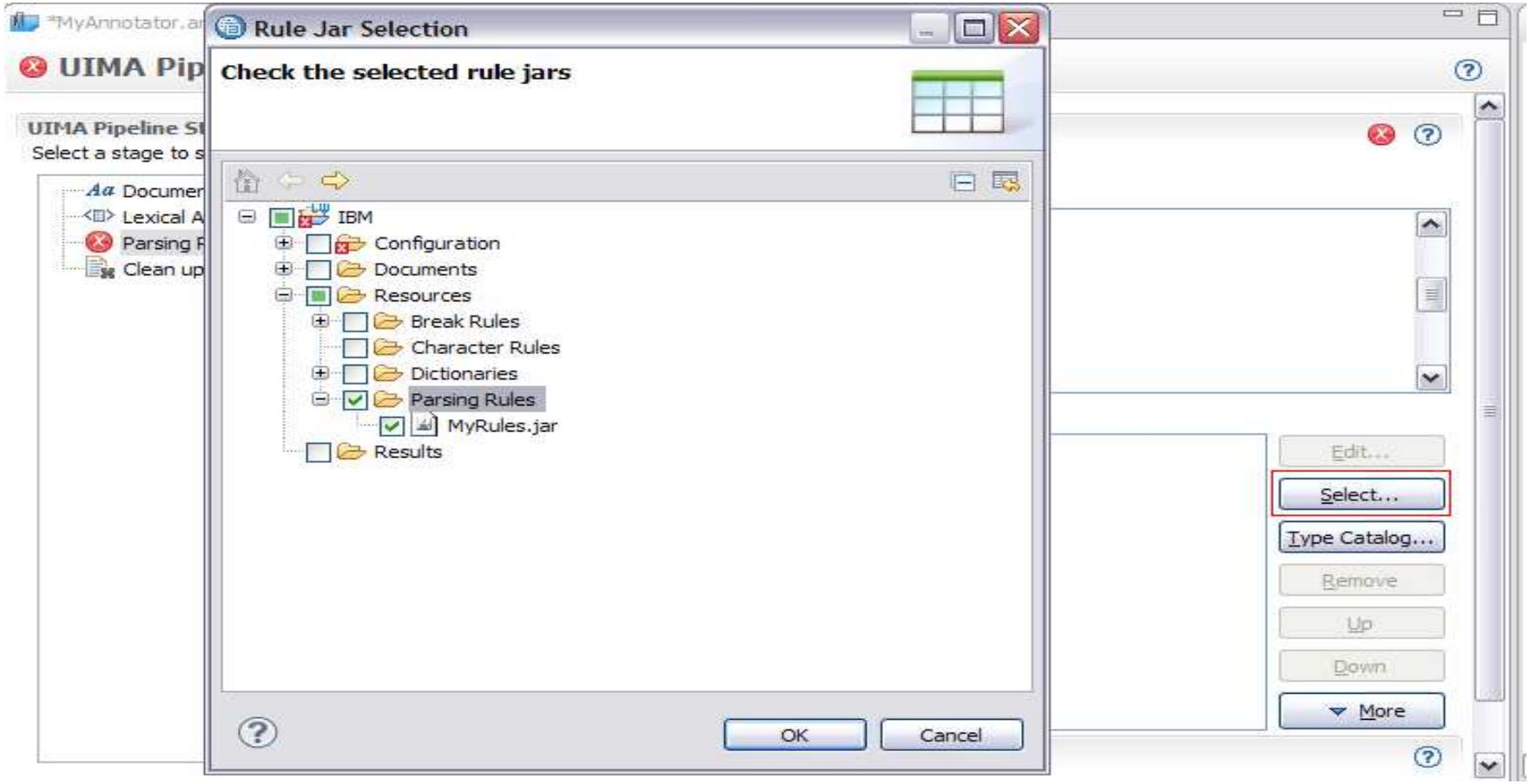


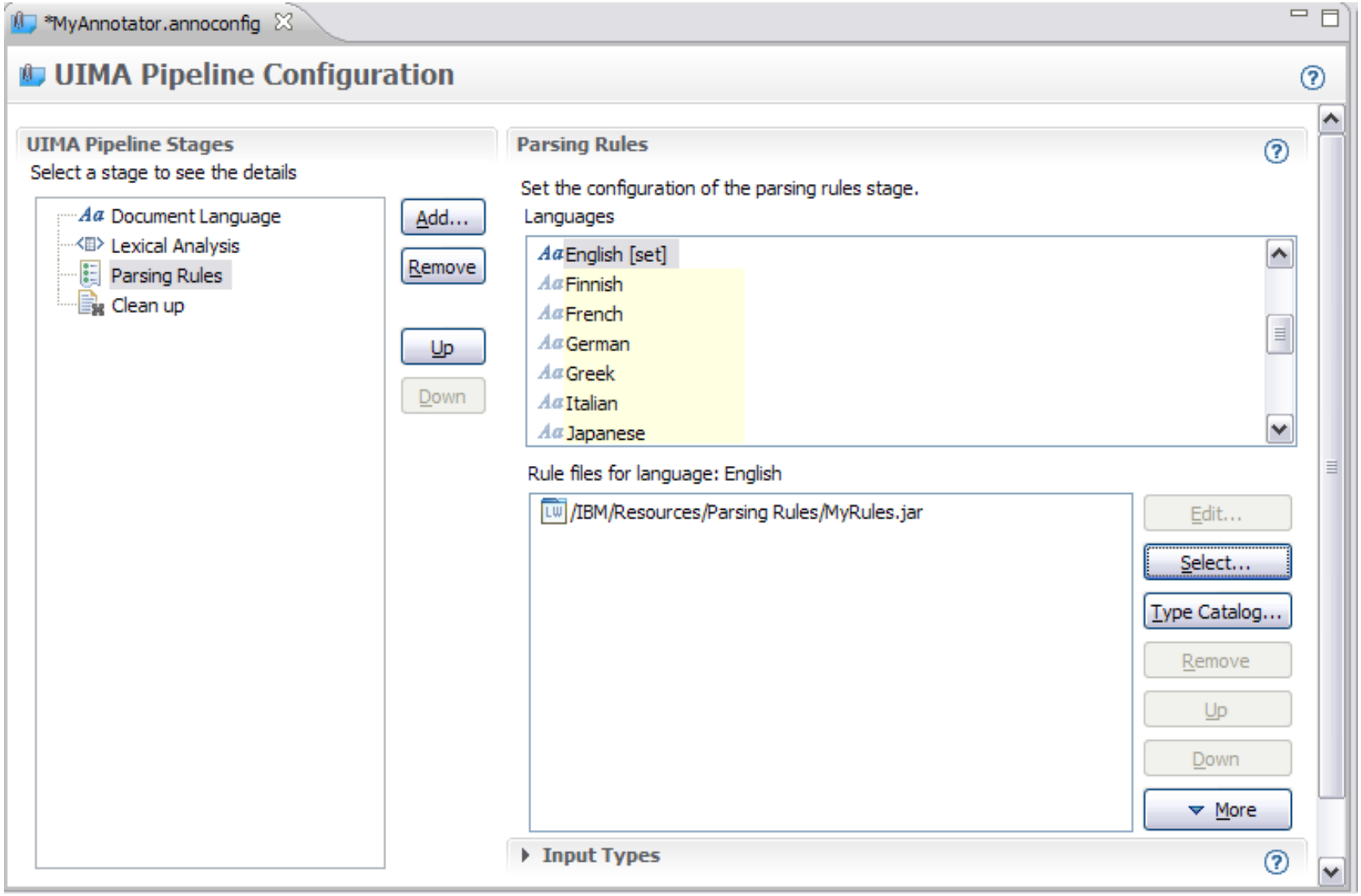
Parsing rules stage

- Make sure the language is set to English



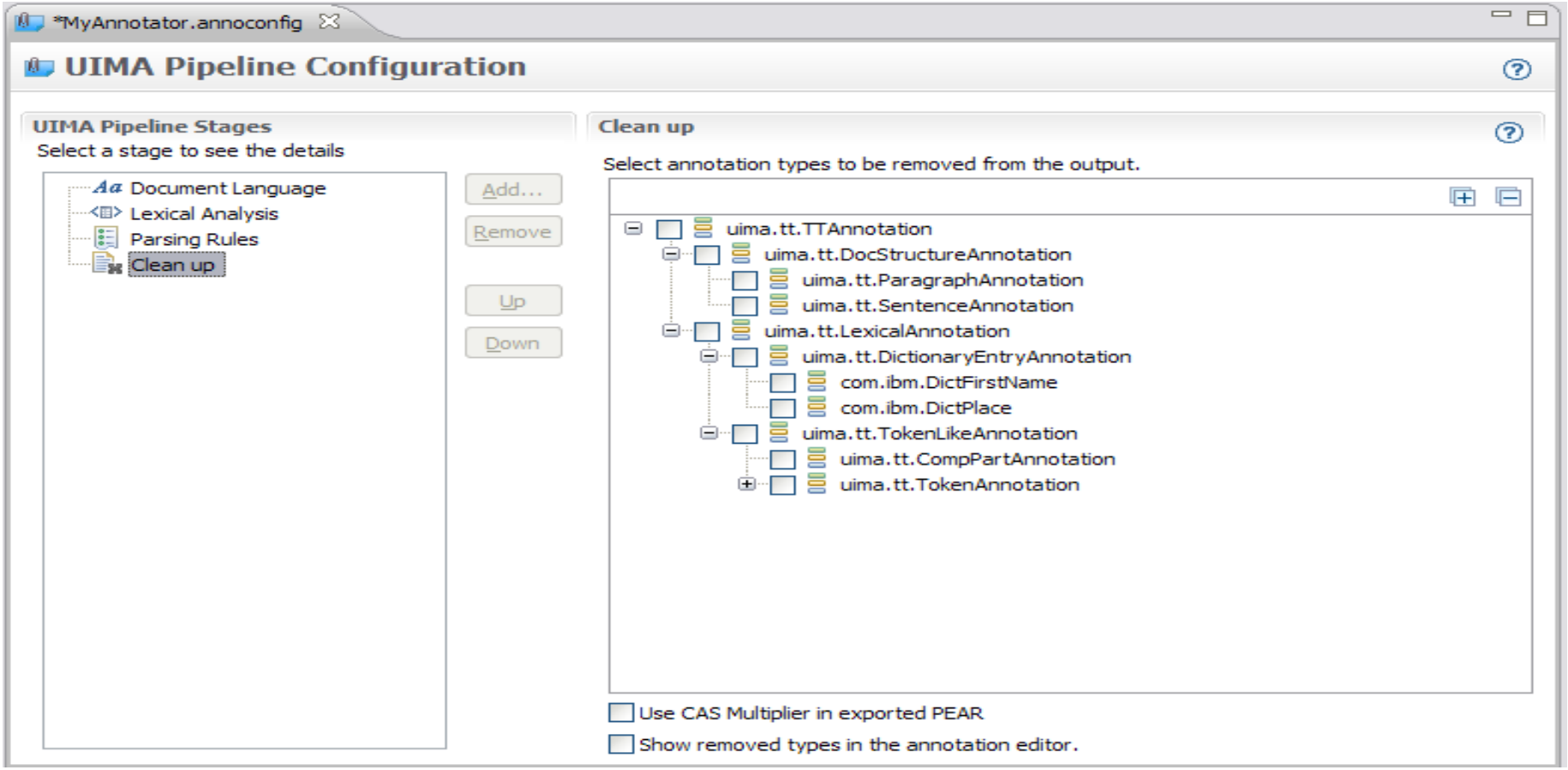
- To add a parsing rules database, click the Select button, this will open the workspace browser allowing you to add the relevant resources to the pipeline. Click OK after you made your selection.
- You can also drag the .jar file from the LanguageWare Explorer into the Parsing Rules field.





Clean up stage

- In this stage, you can select the Types (concepts) that you want to show/hide when annotating documents. This is useful when you have intermediate Types that you don't want to see annotated in the final output of the document analysis. This will be covered in the “Annotate a Document” module.



- After you finish adding all the resources to the UIMA pipeline configuration, save it.
- After this, you can move on to the next tutorial and start annotating documents and see the output.

Module roadmap

- **UIMA pipeline configuration**
  - What is it?
  - How to configure it and run it?
  - How to see the annotation results?
- **Summary and best practices**
- **Sample exercises**

Module summary

You have completed this module and can:

- *Create and configure a UIMA pipeline configuration.*
- *Learn how to annotate documents (add the name of the relevant presentation).*

Refer to the LanguageWare help for more tips and advanced use cases.

Best practices

- The UIMA pipeline configuration is the place holder for the configuration parameters used to annotate documents for a specific model.
- The pipeline contains all the relevant information (document language, output types to be shown/hidden) and resources (lexical and rules) for an annotator.
- More advanced options can be configured in the pipeline; they are covered in the help.
- More resources can be added to the pipeline, so you can add new custom dictionaries and parsing rules databases as needed.
- It's a good practice to give relevant names to the pipeline.

Module roadmap

- **UIMA pipeline configuration**
  - What is it?
  - How to configure it and run it?
  - How to see the annotation results?
- **Summary and best practices**
- **Sample exercises**

Practice exercises

- Create an annotator called "AnalyseHelpline"
- Add "Flavor" Dictionary and "IdentifyQuestion" Parsing Rules to the pipeline

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

IBM, the IBM logo, ibm.com, developerWorks, and LanguageWare are trademarks or registered trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of other IBM trademarks is available on the web at "Copyright and trademark information" at http://www.ibm.com/legal/copytrade.shtml

Other company, product, or service names may be trademarks or service marks of others.

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.