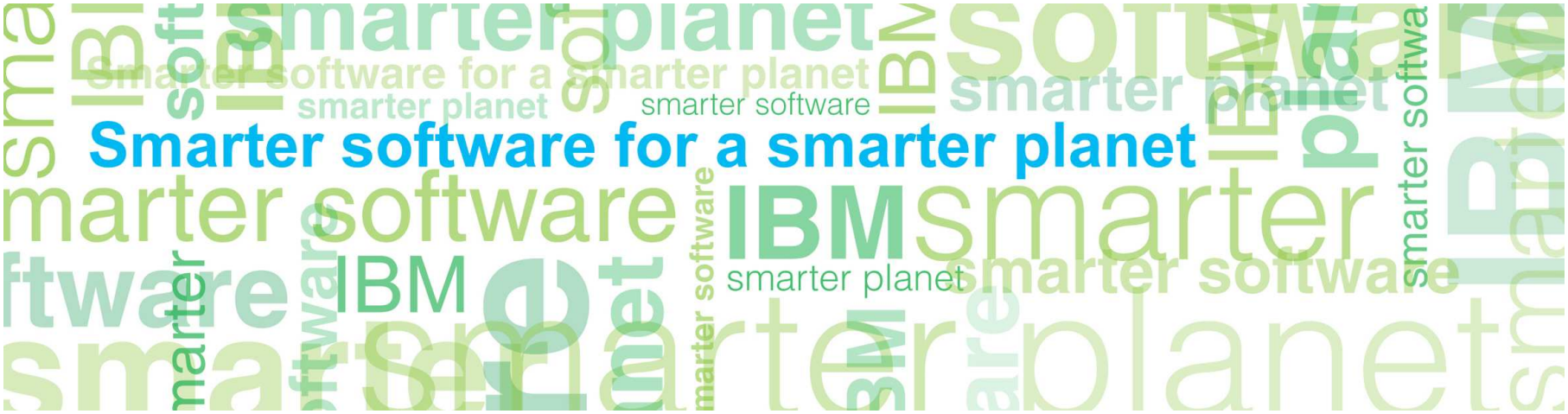
LanguageWare Resource Workbench 7.2

Create and use break rules



Introduction

- **Module overview**
  - What are break rules
  - How to create, configure and use break rules
  - Best practices
- **Target audience:**
  - All audiences
- **Prerequisites:**
  - Install LanguageWare® Resource Workbench
  - Create a LanguageWare project
  - Create a UIMA annotator
  - Annotate a document
- **Version Release Date:** LRW 7.2, ICA 2.2, released October, 2010

Module objectives

After this module you will be able to:

- Understand how break rules affect text tokenization.
- Create a break rules file.
- Configure a break rules file.
- Build a break rules file and use it in your annotator.

Module roadmap

- **Break rules**
  - What are break rules
  - Creating a break rules file
  - Configuring the break rules file
  - Building the break rules file
  - Using the break rules in an annotator
- **Summary and best practices**
- **Sample exercises**

Break rules

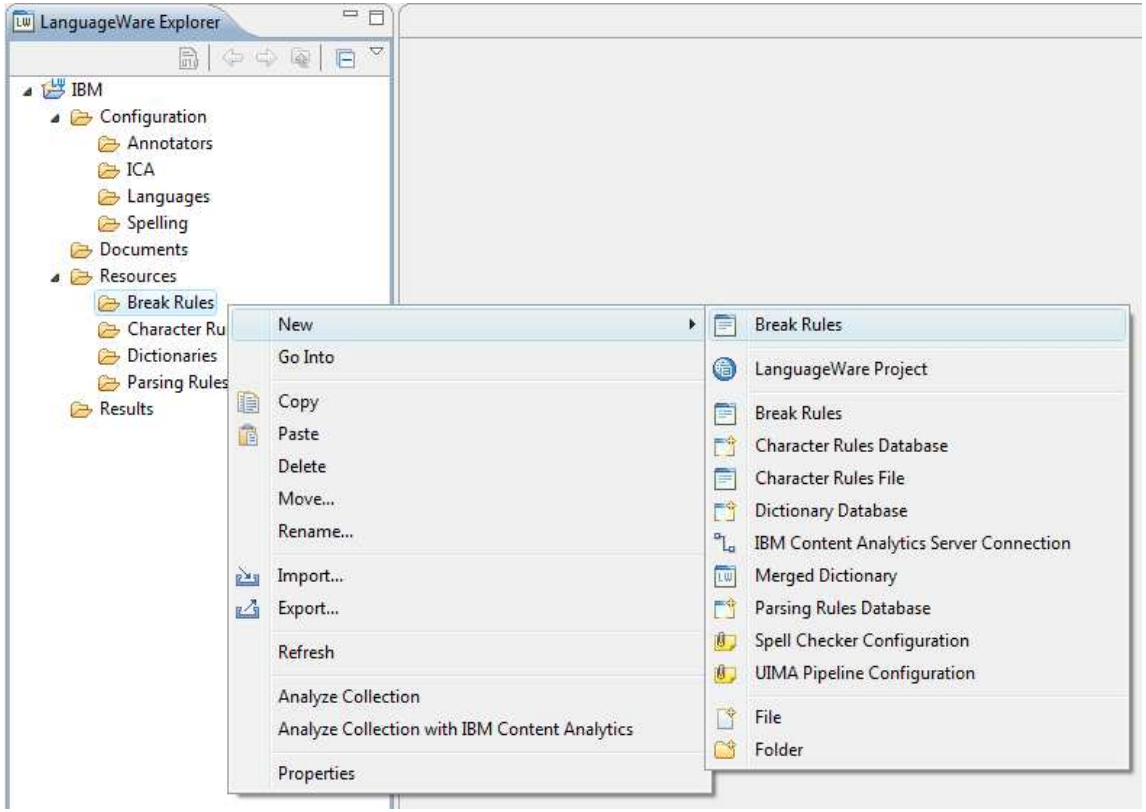
What are they?

- LanguageWare uses break rules to split documents into lexical components.
- These lexical components are paragraphs, sentences and tokens.
- A token is a unit of text, and could be a word, a punctuation symbol, a number, or a string of symbols.
- Much of the logic of how to split a document into components is standard, and does not need to be defined.
- However, there are several options where the document structure, and your preferences, dictate how a document should be split into components.
- For example, in some documents, a new line is an indication of the start of a new paragraph, while in other documents the new line is only a mechanism to format the document into lines of a fixed length. For this reason, LanguageWare allows you to configure these options in the break rules.

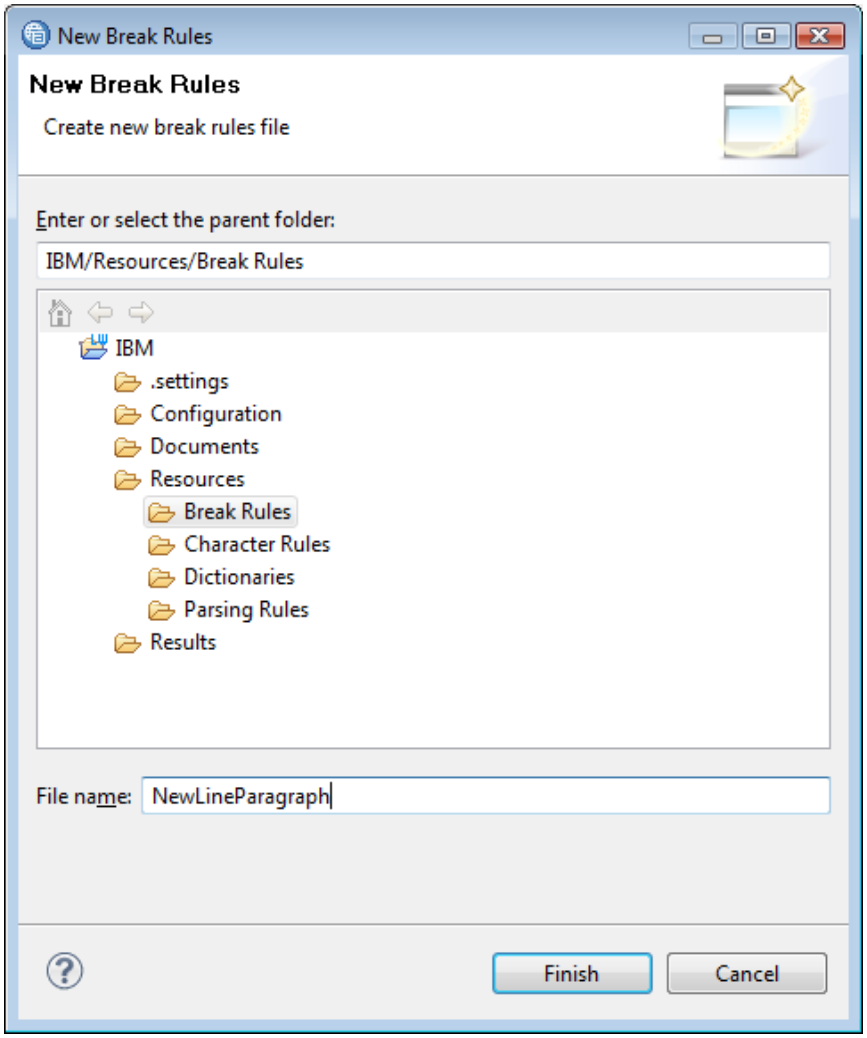
Break rules file

How to create it?

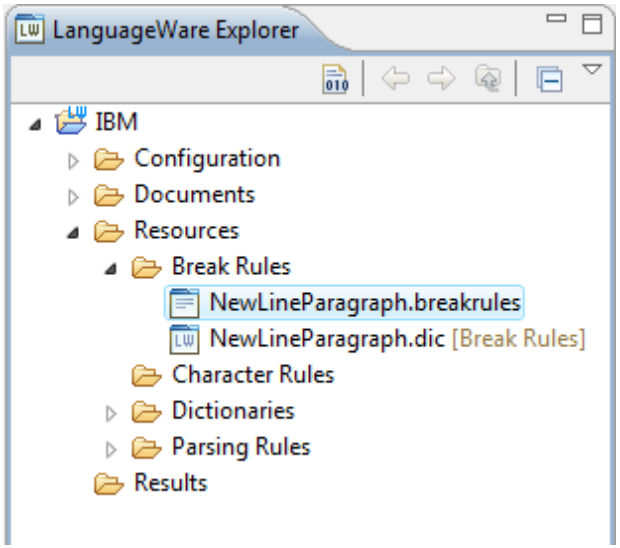
- Right click the break rules folder in the Resources directory, and select New/Break Rules.



- The New Break Rules dialog box pops up:
  - The parent folder will match the one you right clicked (It can be changed if needed).
  - Specify the name of the Break Rules file. It is good to use relevant names (specific to the function, for example NewLineParagraph, or relevant to the model, for example Healthcare).
- Click Finish.

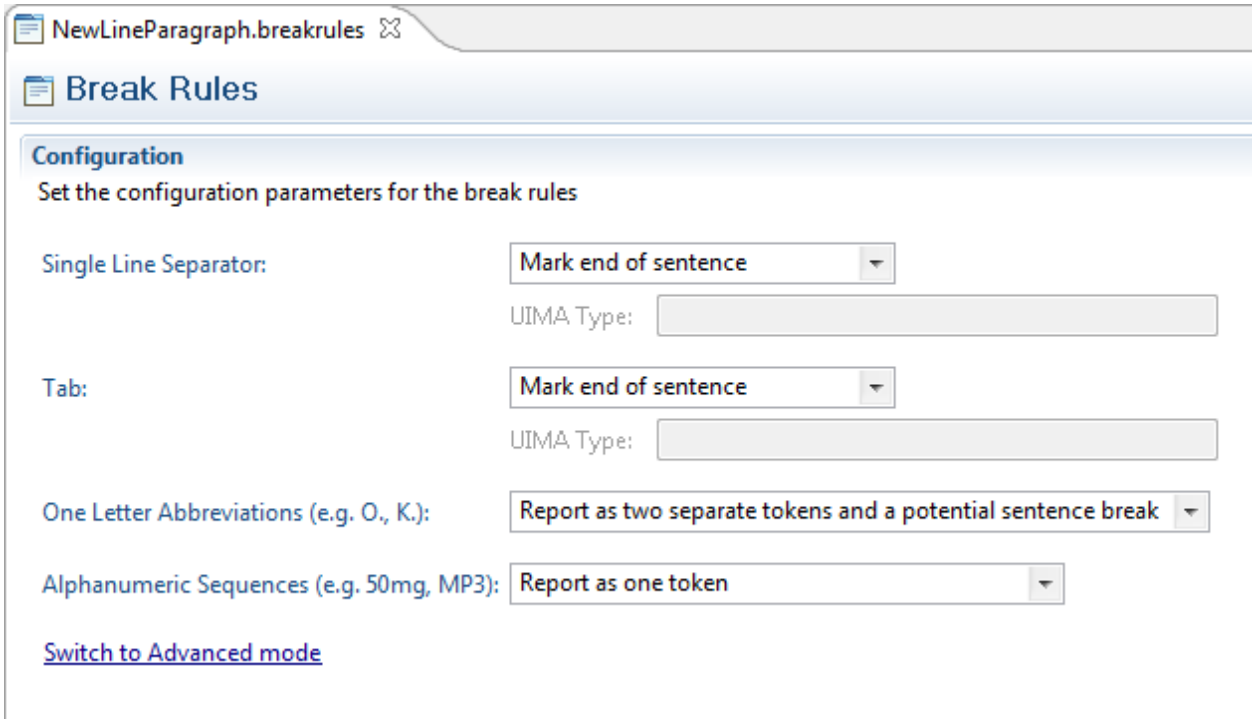


- The following files are generated when creating break rules.
  - **.breakrules**: Contains the source of the break rules. This is the file you edit to change the settings.
  - **.dic**: The break rule dictionary which contains a compiled version of the rules. This dictionary is used when analyzing a document.



Break rules file
Setting options

- When you create a new break rules file, the **.breakrules** file is automatically opened in the break rules editor to allow you to set required options.



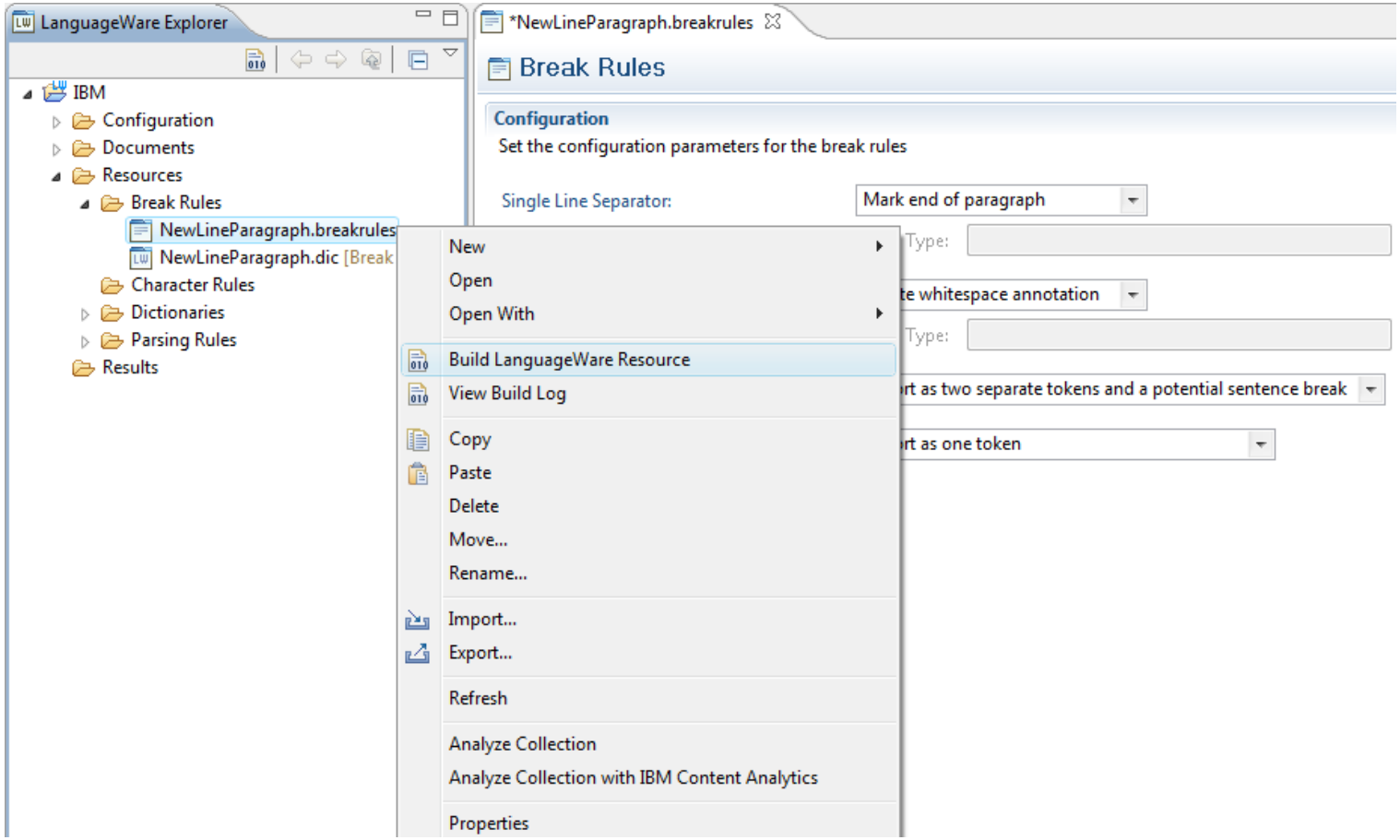
Break rule Options

Single line separator	A single line separator between two lines of text. This can indicate: * End of paragraph * End of sentence * Punctuation * Whitespace
Tab	A tab character. This can indicate: * End of paragraph * End of sentence * Punctuation * Whitespace
One letter abbreviation	A single alphabetic character followed by a period, for example **B.** This can be treated as: * A single token * Two separate tokens and a potential sentence break.
Alphanumeric sequences	A string containing both alphabetic and numeric characters, for example **50mg.** This can be treated as: * A single token * Separate alphabetic and numeric tokens

- Once you have set the options, select the menu options **File / Save** and then **File / Close**.

Break rules file
Building

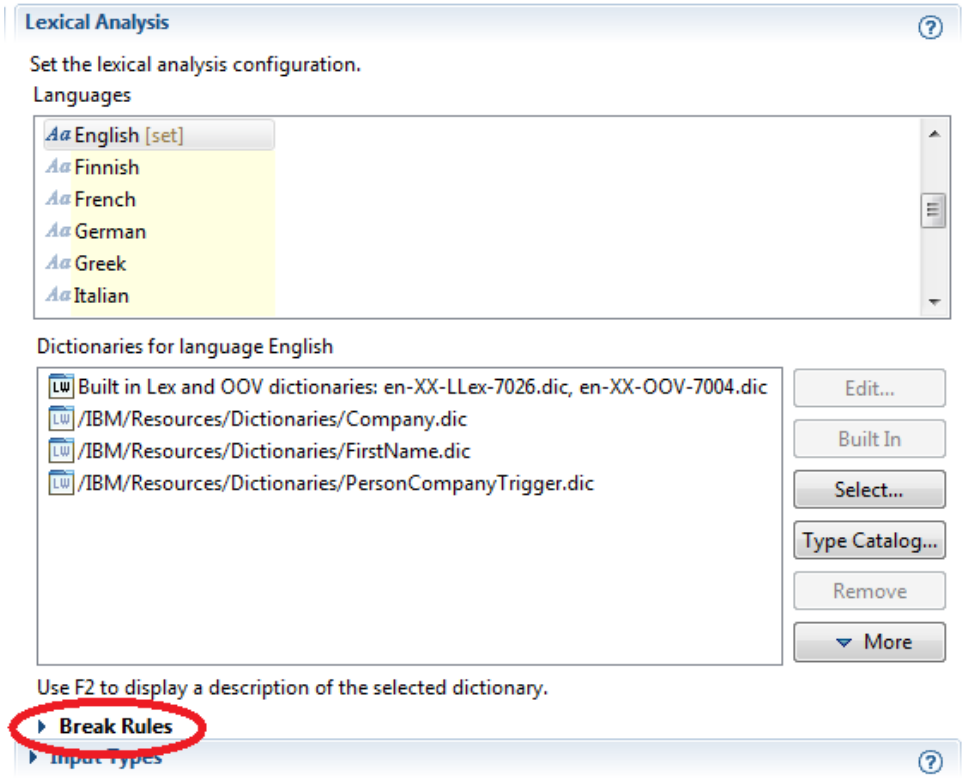
- Once you have updated the **.breakrules** file, right click it and select Build LanguageWare Resource to build the updated rules into the **.dic** file.



Break rules

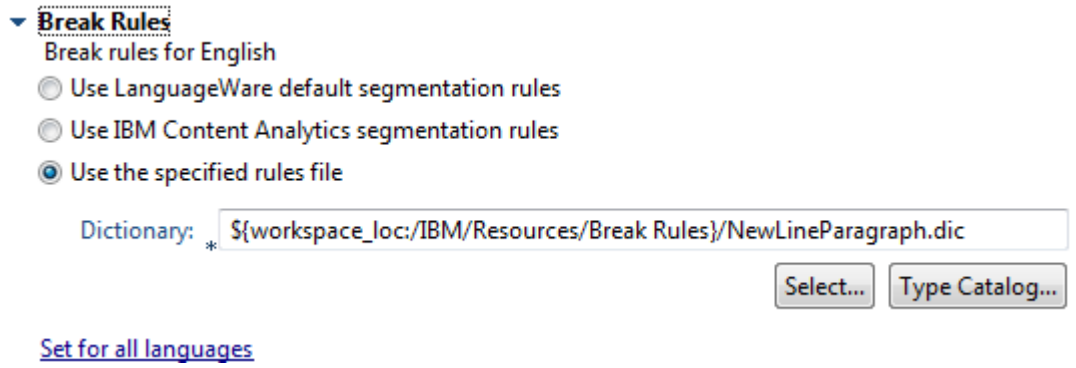
Using in an annotator

- Open a UIMA pipeline configuration file (file type .annoconfig)
- Select the **Lexical Analysis** stage in the list of UIMA pipeline stages.
- Break rules are specified in the **Break Rules** section which is normally collapsed.
- Expand this section by clicking it.



Break rules

Using in an annotator



- Click **Use the specified rules file**
- Click the **Select** button
- Use the **Dictionary Selection** dialog to select the Break Rules dictionary, for example, NewLineParagraph.dic
- If this break rules file should be used for all languages, click the **Set for all languages** link.
- Save the changes to the UIMA pipeline configuration file.

Module roadmap

- **Break rules**
  - What are break rules
  - Creating a break rules file
  - Configuring the break rules file
  - Building the break rules file
  - Using the break rules in an annotator
- **Summary and best practices**
- **Sample exercises**

Module summary

You have completed this course and can:

- Create a break rules file
- Configure and build a break rules file into a dictionary
- Use a break rules dictionary in a UIMA pipeline

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

Best practices

- A break rules file tells the annotator how to tokenize the text while annotating a document.
- Give the break rules file a meaningful name so it is obvious what its purpose is.

Module roadmap

- **Break rules**
  - What are break rules
  - Creating a break rules file
  - Configuring the break rules file
  - Building the break rules file
  - Using the break rules in an annotator
- **Summary and best practices**
- **Sample exercises**

Practice exercises

- Analyze all documents in the documents folder of the "ConfectionaryCompanyHelpline" project with the "AnalyseHelpline" annotator.
- Save the annotations in a file called “DefaultBreak”.
- Create a break rules file called “HelplineBreakrules”
- Set the “Alphanumeric sequences” option to report as separate tokens
- Rebuild the break rules file and include the file in the "AnalyseHelpline" annotator.
- Again, analyze all documents in the documents folder of the "ConfectionaryCompanyHelpline" project with the "AnalyseHelpline" annotator.
- Compare the results of these annotation with the saved annotations in the file “DefaultBreak”.

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.