Homework 1: Text Complexity Measurement Assignment

For this assignment, you are provided with a zip file containing novels from Project Gutenberg (link). Each novel is saved as a `.txt` file. Your task is to measure the complexity of these novels using three different readability indices: the Automated Readability Score (ARS), the Flesch-Kincaid Index (FKI), and the Dale-Chall Readability Score.

1 The Automated Readability Score (ARS)

The Automated Readability Score (ARS) is based on the entire text of each novel. For more information, see this link.

The ARS is calculated using two parameters:

- The average number of characters per word (`cpw`)
- The average number of words per sentence (`wps`)

The formula for ARS is:

$$ARS = 4.71 \cdot cpw + 0.5 \cdot wps - 21.43 \tag{1}$$

A higher ARS value indicates more complex language. The ARS is designed to approximate the US grade level required to understand the text. For example, an ARS of 12 corresponds roughly to a high school senior reading level.

2 Flesch–Kincaid Index (FKI)

The Flesch-Kincaid Index (FKI) uses the average number of words per sentence (wps) and the average number of syllables per word (spw). For more details, see this link.

The formula for FKI is:

$$FKI = 0.39 \cdot wps + 11.8 \cdot spw - 15.59 \tag{2}$$

Similar to the ARS, a higher FKI value indicates a more complex text. The FKI is used by the US military and is also designed to approximate the US grade level needed to comprehend the text.

3 Dale-Chall Readability Score

The Dale-Chall Readability Score is based on the percentage of difficult words in a text. A word is considered difficult if it has more than 2 syllables. For more information, see this link.

The formula for the Dale-Chall Readability Score is:

$$\text{Dale-Chall} = 0.1579 \cdot \left(\frac{\text{difficult\_words}}{words} \cdot 100\right) + 0.0496 \cdot wps \tag{3}$$

The Dale-Chall Readability Score corresponds to the following US grade levels:

- 4.9 and below: Grade 4 and below
- 5.0 to 5.9: Grades 5-6
- 6.0 to 6.9: Grades 7-8
- 7.0 to 7.9: Grades 9-10
- 8.0 to 8.9: Grades 11-12
- 9.0 to 9.9: Grades 13-15 (College)
- 10 and above: Grades 16 and above (College Graduate)

4 Assignment Tasks

Complete the following steps:

1. Extract the files from the zip archive into a folder named 'Novels'.
2. Use the `os` library to obtain the path of each novel.
3. Read the text content of each novel. A novel is any file with the extension `.txt`.
4. Process the text, removing punctuation where necessary.
5. Calculate the ARS, FKI, and Dale-Chall score for each novel.
6. Display the novel name, ARS, FKI, and Dale-Chall scores (rounded to 3 significant figures).

You are required to create a function that calculates the ARS, FKI, and Dale-Chall score for a given text.

```
def measure_complexity(filepath):
    # Function implementation here
```

5 Implementation Guidelines/Hints

Follow these steps to implement your solution:

1. Divide the text into a list of sentences. For simplicity, consider the period (.) as the sole sentence delimiter. (Substitute all ! with .)
2. Count the total number of sentences and store this value.
3. Use a loop to iterate over each sentence, splitting it into a list of words.
4. Implement a nested loop to process each word, counting its characters and syllables.
5. Keep track of the total word count, character count, and syllable count.
6. For the Dale-Chall score, count the number of difficult words.

```
def count_syllables(word):
    # Function implementation here

def is_difficult_word(word):
    # Function implementation here
```

The `count\_syllables` function should take a string (word) as input and return an integer representing the number of syllables in that word.

For this assignment, use the following simplified method to count syllables:

- Remove any trailing 's' or 'e' from the word.
- Count the transitions between consonants and vowels (a, e, i, o, u) in the word.

You can use `count\_syllables` within your `is\_difficult\_word` function. It takes a word as an input, and return a `1` or `0` indicating whether the word is considered difficult.

For example:

```
>>> count_syllables('creeps')
1
>>> count_syllables('devotion')
3
>>> is_difficult_word('cat')
0
>>> is_difficult_word('ubiquitous')
1
```

Once you have calculated the necessary values (word count, sentence count, character count, syllable count, and difficult word count), you can compute the ARS, FKI, and Dale-Chall score for any given text.

Your final function should accept a file's path as input and output the file name, ARS, FKI, and Dale-Chall score.