Towards Arabic Text Simplification

Nouran Khallaf, mlnak@leeds.ac.uk
School of Languages, Cultures and Societies, University of Leeds, UK

A system to simplify Arabic texts by reducing the lexical and syntactic complexity using a hybrid method combining Machine Learning and Rule-Based techniques

1. **Text Simplification**
   - **Definition**: Simplifying a text is the process of reducing its linguistic complexity, while maintaining its meaning and original information.

2. **Importance**
   - **Usage in designing and simplifying the language curriculum for both second language and first language learners**
   - **Make text easy-to-read for first language users with cognitive impairments and low literacy language level**

3. **Challenges**
   - Highly morphologically rich language
   - Flexible word order
   - Multifunctionality of Arabic nouns
   - Lack of vocalisation diacritics
   - Lack of Arabic resources: Datasets, Corpora, Arabic NLP tools

4. **Assessing Readability**
   - **Stage I**: To measure and annotate the difficulty of the Arabic text.
   - **Stage II**: Adopting CEFR [A1, A2, B3, B4, C1, C2]
   - **Based on average of complex lexical items and complex sentence structure.**

5. **Lexical Simplification**
   - **Not all the words in the text need to be simplified**

6. **Syntactic Simplification**
   - **Syntactic Dependency Parsing**
   - **Identify complex structure**
   - **Transformation rules**
   - **Generation/Regeneration**

7. **New Arabic Frequency List**
   - New classified Arabic frequency list consisting of 8834 unique Lemmas from Buckwalter, Al-kitaab and KELLY's

Example

|----|----|----|----|----|----|----|----|----|----|----|----|----|

The Parliament should send a message, because that is the majority of people want.