## Towards Arabic Text Simplification UNIVERSITY OF LEEDS

Nouran Khallaf, mlnak@leeds.ac.uk School of Languages, Cultures and Societies, University of Leeds, UK **Supervised by**: Dr. Serge Sharoff Prof. Michael Ingleby

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

Why?

**Text Simplification** 

**Definition** 

**Complex** 

Importance

Usage in designing and simplifying the

Simplifying a text is the process of reducing its linguistic complexity, while maintaining its meaning and original information.

A man carrying a large number of books entered the room

دخل رجل يحمل عدد كبير من الكتب إلى الغرفق

دخل رجل الي الغرف مع الرجل كتب كثير

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

A man walked into the room. The man had many books.

One Complex sentence transformed to two simple ones

A fundamental pre-process in NLP applications such as text retrieval, extraction, summarization, translation

Challenges

Simple

Highly morphologically rich language

Flexible word order

Multifunctionality of Arabic nouns

Lack of vocalisation diacritics

Lack of Arabic resources: DatasetsCorporaArabic NLP tools **Assessing Readability** 

To measure and annotate the difficulty of the Arabic text.

Adopting CEFR [A1,A2,B3,B4,C1,C2]

Based on average of complex lexical items and complex sentence structure.

**Lexical Simplification** 

Not all the words in the text needs to be simplified

Identify the Complex Word

Generate series of substitutions

Substitution Ranking based on context

Select the new simple Substitute

**Example Complex Grammatical structure** 

A1 A2 B2

**B2** 

لذا ينبغي على البرلمان أن يبعث برسالة لأن هذه هي رغبة الأغلبية الساحقة The Parliament should send a message, because that is the majority of people want.

Assign Level for each word

**C2** 

Identify complex grammatical structure

Assign Readability Level for Sentence

**Syntactic Simplification** 

Syntactic Dependency Parsing

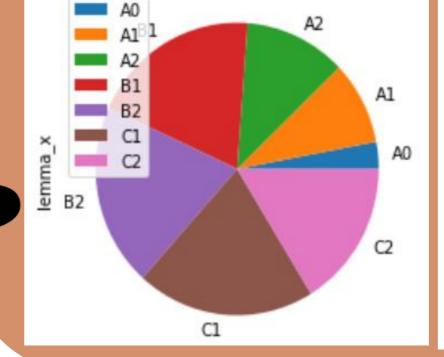
Identify complex structure

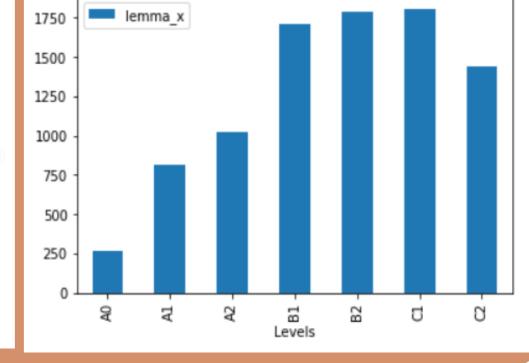
Transformation rules

Generation/ Regeneration

New Arabic Frequency List

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







Stage III





