

# The SAMER Arabic Text Simplification Corpus

**Bashar Alhafni, Reem Hazim, Juan Piñeros Liberato,  
Muhammed Al Khalil, Nizar Habash**

# Motivation and Related Work

نحن يا سيدتي في زمن مضطرب لا يركد عجاجه، ولا تسكن سيوفه في أغمادها، بعد أن انحلت أواصر بني العباس، وأصبحت دولتهم أشلاء ممزقة، يفترسها كل مفترس، ويُغير عليها كلّ واثب.

We live, my lady, in a troubled age whose dirt-laden air does not settle, nor its swords rest in their scabbards, after the bonds of the house of Abbas dissolved, and their state became like torn limbs preyed upon by every predator and raided by every opportunist.

# Motivation and Related Work

نحن يا سيدتي في زمن مضطرب لا يركد عجاجة، ولا تسكن سيوفه في أغمادها، بعد أن انحلت أواصر بني العباس، وأصبحت دولتهم أشلاء ممزقة، يفترسها كل مفترس، ويُغير عليها كل واثب.

*We live, my lady, in a troubled age whose dirt-laden air does not settle, nor its swords rest in their scabbards, after the bonds of the house of Abbas dissolved, and their state became like torn limbs preyed upon by every predator and raided by every opportunist.*

# Motivation and Related Work

نحن يا سيدتي في زمن مضطرب لا يركد عجاجه، ولا تسكن سيوفه في أغمادها، بعد أن انحلت أواصر بني العباس، وأصبحت دولتهم أشلاء ممزقة، يفترسها كل مفترس، ويُغير عليها كل واثب.

*We live, my lady, in a troubled age whose dirt-laden air does not settle, nor its swords rest in their scabbards, after the bonds of the house of Abbas dissolved, and their state became like torn limbs preyed upon by every predator and raided by every opportunist.*



نحن يا سيدتي في زمن صعب لا يهدأ غباره، ولا تسكن سيوفه في جيوبها، بعد أن تقطعت صلات بني العباس، وأصبحت دولتهم قطعاً ممزقة، يفترسها كل مفترس، ويهجم عليها كل عدو.

*We live, my lady, in a difficult age whose dust does not clear and whose swords do no rest in their sheaths, after the joints of the house of Abbas were cut off, and their state became like torn body parts preyed upon by every predator and attacked by every enemy.*

# Motivation and Related Work

## Text Simplification Approaches:

- **Lexical Simplification**
  - Identifying and replacing difficult words with easier synonyms while preserving the information and meaning of the original text
- **Syntactic Simplification**
  - Making complex syntactic constructions simpler. For example, replacing passive voice constructions with active voice
- **End-to-End Simplification**
  - Both lexical and syntactic simplification. Usually defined as a sequence-to-sequence task

# Motivation and Related Work

## Text Simplification Datasets:

- Mostly in English: LexMTurk (Horn et al., 2014), LSeval (De Belder and Moens, 2012), Newsela Corpus (Xu et al. 2015), BenchLS (Paetzold and Specia, 2016), TurkCorpus (Xu et al. 2016)
- No publicly available Arabic text simplification datasets

# Motivation and Related Work

## Text Simplification Datasets:

- Mostly in English: LexMTurk (Horn et al., 2014), LSeval (De Belder and Moens, 2012), Newsela Corpus (Xu et al. 2015), BenchLS (Paetzold and Specia, 2016), TurkCorpus (Xu et al. 2016)
- No publicly available Arabic text simplification datasets

## In this work:

- We create the first manually annotated Arabic dataset for **lexical simplification**
- 159K words selected from 15 Arabic fiction novels
- Readability level annotations at both the document and word levels
- Two simplification parallels for each text targeting school-aged learners at two different readability levels

# Roadmap

- Motivation
- **Arabic Linguistic Facts**
- The SAMER Project
- The SAMER Simplification Corpus
- Conclusion



# Arabic Linguistics Facts

- **Morphological richness:**

- Inflects for gender, number, person, case, state, aspect, mood and voice
- 22,400 possible POS tags (48 POS tags for English)
- ~3 different core meanings (i.e., lemmas) per word

- **Orthographic ambiguity:**

- Optional diacritics to specify short vowels and consonantal doubling
- ~7 diacritizations per word

# Arabic Linguistics Facts

- **Morphological richness:**

- Inflects for gender, number, person, case, state, aspect, mood and voice
- 22,400 possible POS tags (48 POS tags for English)
- ~3 different core meanings (i.e., lemmas) per word

- **Orthographic ambiguity:**

- Optional diacritics to specify short vowels and consonantal doubling
- ~7 diacritizations per word

درסהا

دَرْسُهَا	دَرَّسَهَا	دَرَسَهَا
<i>darsu+hA</i>	<i>dar~asa+hA</i>	<i>darasa+hA</i>
<i>her lesson</i>	<i>he taught her</i>	<i>he studied it</i>

# Arabic Linguistics Facts

- **Morphological richness:**

- Inflects for gender, number, person, case, state, aspect, mood and voice
- 22,400 possible POS tags (48 POS tags for English)
- ~3 different core meanings (i.e., lemmas) per word

- **Orthographic ambiguity:**

- Optional diacritics to specify short vowels and consonantal doubling
- ~7 diacritizations per word

درסהا

دَرَسُ + ها      دَرَّسَ + ها      دَرَسَ + ها

*darsu+hA*      *dar~asa+hA*      *darasa+hA*

*her lesson*      *he taught her*      *he studied it*

# Roadmap

- Motivation
- Arabic Linguistic Facts
- **The SAMER Project**
- The SAMER Simplification Corpus
- Conclusion

# The SAMER Project

- Simplification of **Arabic Masterpieces for Extensive Reading (SAMER)**
  - Funded by New York University Abu Dhabi Research Enhancement Fund
- **Objective:** to create standards and tools for the simplification of Arabic fiction to school-aged learners
- The project has developed the following open-source resources:
  - **Readability Lexicon:** A five-level readability scale at the word-level
  - **Readability Thesaurus:** An Arabic Readability-leveled Thesaurus for Arabic
  - **Simplification Interface:** A simplification interface platform as a Google Docs Add-on
  - **Simplification Corpus:** A text simplification corpus of Arabic novels targeting two readability levels

# Roadmap

- Motivation
- Arabic Linguistic Facts
- The SAMER Project
- **The SAMER Simplification Corpus**
- Conclusion

# SAMER Simplification Corpus

## Corpus Selection:

- 15 publicly available Arabic fiction novels from the Hindawi foundation
- 14 novels were published between 1865 and 1955. One famous novel published in the 12th century
- Extracted the first ~10K words from each novel (~159K words in total)
- Segmented chapters into paragraphs of 1500 words (4289 paragraphs)

Book Title	Author	Date	Para.	Words
حي بن يقظان <i>Hayy Ibn Yaqzan</i>	ابن طفيل <i>Ibn Tufail</i>	1150	213	9,962
غابة الحق <i>The Forest of Truth</i>	فرانسيس مراث <i>Francis Marrash</i>	1865	241	10,081
لادياس <i>Ladiyas</i>	أحمد شوقي <i>Ahmed Shawqi</i>	1899	300	9,846
المحالفة الثلاثية في المملكة الحيوانية <i>The Tripartite Alliance of the Animal Kingdom</i>	أمين ربحاني <i>Ameen Rihani</i>	1904	315	10,577
الملك كورش <i>Cyrus the Great</i>	زينب فواز <i>Zaynab Fawwaz</i>	1905	235	9,910
الأجنحة المتكسرة <i>Broken Wings</i>	جبران خليل جبران <i>Kahlil Gibran</i>	1912	170	11,482
زينب <i>Zaynab</i>	محمد حسين هيكل <i>Mohammed Hussein Heikal</i>	1913	138	9,861
شجرة الدر <i>The Pearl Tree</i>	جرجي زيدان <i>Jurji Zaydan</i>	1914	296	12,230
إبراهيم الكاتب <i>Ibrahim Al-Katib</i>	إبراهيم عبد القادر المازني <i>Ibrahim Abd Al-Qadir Al-Mazini</i>	1931	295	10,173
ثورة في جهنم <i>A Revolution in Hell</i>	نقولا حداد <i>Niqula Haddad</i>	1938	492	11,713
سارة <i>Sara</i>	عباس محمود العقاد <i>Abbas Mahmoud Al-Aqqad</i>	1938	300	10,079
فارس بني حمدان <i>The Knight of Beni Hamdan</i>	علي الجارم <i>Ali Al-Jarem</i>	1945	329	10,820
على باب زويلة <i>On Bab Ziwaita</i>	محمد سعيد العريان <i>Mohammed Saeed Al-Aryan</i>	1951	313	11,701
نماذج بشرية <i>Human Examples</i>	أحمد رضا حوحو <i>Ahmad Rida Huhu</i>	1955	443	10,080
هذا التاج <i>This Crown</i>	واصف البارودي <i>Wasef Al-Baroudi</i>	1955	209	10,750
			<b>4,289</b>	<b>159,265</b>

# SAMER Simplification Corpus

## Corpus Annotation:

- **Goal:** simplify the Arabic fiction novels so that they can be targeted towards school-aged learners
- We use the five readability levels that were defined by Al Khalil et al. (2020) when creating the **SAMER Lexicon** (40K lemmas)



# SAMER Simplification Corpus

## Corpus Annotation:

- **Goal:** simplify the Arabic fiction novels so that they can be targeted towards school-aged learners
- We use the five readability levels that were defined by Al Khalil et al. (2020) when creating the **SAMER Lexicon** (40K lemmas)

Level	Grade	Age	Examples
Level I	Grade 1	6	بَيْت، شَجَرَة، أُرْنَب، أُرْزَق، كَبِير، صَنَعَ، أَكَلَ، فَرَحَ، عَلَى، لَكِنْ house, tree, rabbit, blue, big, to make, to eat, to be happy, on, but
Level II	Grades 2-3	7-8	جَزِيرَة، ذَهَب، سَنَة، دَاكِن، أَشْطَوَانِي، صَغْب، خَدَعَ، كَافَأً، قُرْبَ، إِذَا island, gold, year, dark, cylindrical, difficult, to cheat, to reward, near, if
Level III	Grade 4-5	9-10	رَيْة، مُتَحَف، مُعَادَلَة، مُمَكِّن، مُوَحَّد، أَغْرَى، نَذَرَ، لَدَى، كَيْ، مَا إِنْ...حَتَّى lung, museum, equation, possible, united, to entice, to be rare, with, for, no sooner... than...
Level IV	Grades 6-8	11-14	اِقْتِصَاد، نُسْغ، طَمَأْنِينَة، رَاقِي، مُثَبَّت، نَكَّتْ، أَغْضَى، إِبَان، إِيْمَا، نُنْ economy, sap, tranquility, sophisticated, proven, to breach, to overlook, during, whereas, if (were)
Level V	Specialist	15 -	أَدَمَة، قَشْطَرَة، هَيْضَة، وَطْيَاف، لَوْذَع، شُعْبِي، لَحَا، خَطَرَن، لَدُنْ، أُنَى epidermis, catheterization, cholera, spectroscope, witty, bronchial, to denounce, to depart, with ( $\approx$ chez in French), wherever

# SAMER Simplification Corpus

## Corpus Annotation:

- **Goal:** simplify the Arabic fiction novels so that they can be targeted towards school-aged learners
- We use the five readability levels that were defined by Al Khalil et al. (2020) when creating the **SAMER Lexicon** (40K lemmas)
- We consider the document readability level to be equal to the highest readability level found among the words in the document
- Given the nature of the documents, all of them will have a readability of **Level 5**
- We focus on producing two simplified versions of each document: **Level 4** and **Level 3**

Level	Grade	Age	Examples
Level I	Grade 1	6	بَيْت، شَجَرَة، أُرْزَق، كَبِير، صَنَعَ، أَكَلَ، فَرَحَ، عَلَى، لَكِنْ house, tree, rabbit, blue, big, to make, to eat, to be happy, on, but
Level II	Grades 2-3	7-8	جَزِيرَة، ذَهَب، سَنَة، دَاكِن، أَشْطَوَانِي، صَغْب، حَدَّعَ، كَافَأَ، قُرُوبَ، إِذَا island, gold, year, dark, cylindrical, difficult, to cheat, to reward, near, if
Level III	Grade 4-5	9-10	رَيْة، مُتَحَف، مُعَادَلَة، مُمَكِّن، مُوَحَّد، أَغْرَى، نَذَرَ، لَدَى، كَيْ، مَا إِنْ...حَتَّى lung, museum, equation, possible, united, to entice, to be rare, with, for, no sooner... than...
Level IV	Grades 6-8	11-14	اِقْتِصَاد، نُسْغ، طَمَأْنِينَة، رَاقِي، مُثَبَّت، نَكَتَ، أَغْضَى، إِبْتَان، إِنَّمَا، نُنْ economy, sap, tranquility, sophisticated, proven, to breach, to overlook, during, whereas, if (were)
Level V	Specialist	15 -	أَدَمَة، قَشْطَرَة، هَيْضَة، وَطْيَاف، لَوْذَع، شُعْبِي، لَحَا، خَطَرَن، لَدُنْ، أُنَى epidermis, catheterization, cholera, spectroscope, witty, bronchial, to denounce, to depart, with ( $\approx$ chez in French), wherever

# SAMER Simplification Corpus

## Corpus Annotation:

- **SAMER Simplification Add-on**  
(Hazim et al. (2020))
- All documents were automatically labeled with their word-level readability by using a Python-API version of the add-on:
  - 1) Morphological disambiguation;
  - 2) Lemma and POS lookup in the SAMER lexicon
- Chapters were loaded into 146 Google Docs

The screenshot displays the SAMER Readability Analysis interface. The main area shows Arabic text with various words highlighted in different colors (blue, green, yellow, red) to indicate readability levels. The text is:   
#٠# بالله #١# عليك #١# لا #٢# تطيلي #١# يا #١# ليلي  
#٢# فإن #١# مما #٣# يشير #٤# شجون #١# النفس،  
#١# ويزيد #١# في #٢# ألم #١# الحزين، #٢# أن  
#١# يدفع #١# إلى #٣# العزاء #١# والصبر #١# بكلمات  
#٤# خاوية #٤# متخاذة #٣# حفظها #١# الناس  
#٣# لينثروها #١# في #١# كل #٥# مآتم. #٢# إن  
#١# كل #١# كلمة #١# من #١# هذه #١# يا #١# ليلي  
#٣# شعلة #٤# تخرج #٣# وجدي، #٥# وتضطرم #١# في  
#٣# فؤادي، #٢# إن #١# الحزن #٠# حرم #٠# قدسي  
#٣# يجب #٢# أن #٣# تخشع #١# أمامه #١# الرؤوس  
#٢# بالصمت #٦# والإطراق.

The sidebar on the right contains the following information:

- SAMER Readability Analysis**
- Analyze Readability**
- Buttons: Doc Level, Word Level, Clear
- Modify Markup** (Show, Hide, Minimize, Delete)
- Success! Loading... This may take a moment
- Doc Level Analysis**
- Current Doc %
- Names 5.7%
- Level 1 49.1%
- Level 2 15.1%
- Level 3 17%
- Level 4 7.5%
- Level 5 3.8%
- Unknown 1.9%
- Total word count: 53
- Target Level 5 (checked)
- At Level 5 | 3 words | 5.7%
- Below Level 5 | 50 words | 94.3%
- Buttons: By Token, By Type

# SAMER Simplification Corpus

## Corpus Annotation:

- Word highlights with different colors according to their readability levels
- Word-level markup (#<level>#)
- The add-on employs two additional readability levels: **Level 0** (proper nouns), **Level 6** (unknown words)

The screenshot displays the SAMER Readability Analysis tool. The main window shows Arabic text with words highlighted in different colors (blue, green, yellow, red) and marked with readability levels (e.g., #1, #2, #3, #4, #5, #6). The right sidebar contains the following sections:

- Analyze Readability**: Buttons for Doc Level, Word Level, and Clear.
- Modify Markup**: Buttons for Show, Hide, Minimize, and Delete.
- Success!**: A green status message.
- Loading...**: A message indicating the analysis is in progress.
- Doc Level Analysis**: A table showing the distribution of readability levels across the document.

Level	Percentage
Names	5.7%
Level 1	49.1%
Level 2	15.1%
Level 3	17%
Level 4	7.5%
Level 5	3.8%
Unknown	1.9%

Total word count: 53

Target Level 5 (checked)

At Level 5 | 3 words | 5.7%

Below Level 5 | 50 words | 94.3%

By Token (selected) | By Type

# SAMER Simplification Corpus

## Corpus Annotation:

- Word highlights with different colors according to their readability levels
- Word-level markup (#<level>#)
- The add-on employs two additional readability levels: **Level 0** (proper nouns), **Level 6** (unknown words)
- Support word substitution by displaying suggestions of related words from the Arabic WordNet (Black et al., 2006) (e.g., synonyms, hypernyms, and hyponyms)

The screenshot displays the SAMER Readability Analysis interface. The main text area shows Arabic text with words highlighted in different colors corresponding to readability levels. The text is:   
#0# بالله #1# عليك #1# لا #2# تطلي #1# يا #1# ليلي  
#2# فإن #1# مما #3# يثير #4# شجون #1# النفس،  
#1# ويزيد #1# في #2# ألم #1# الحزين، #2# أن  
#1# يدفع #1# إلى #3# العزاء #1# والصبر #1# بكلمات  
#4# خاوية #4# متخاذة #3# حفظها #1# الناس  
#3# لينثروها #1# في #1# كل #5# ماتم. #2# إن  
#1# كاهن #1# كلمة #1# من #1# هذه #1# يا #1# ليلي  
#3# شعلة #4# تخرج #3# وجدي، #5# وتضبطم #1# في  
#3# فؤادي، #2# إن #1# الحزن #0# حرم #0# قدسي  
#3# يجب #2# أن #3# تخشع #1# أمامه #1# الرؤوس  
#2# بالصمت #6# والإطراق.

The sidebar on the right, titled "SAMER Readability Analysis", shows a table with levels 1 through 5, each with "Assign" and "Assign all" buttons. Below the table, there are sections for "Synonyms", "Hypernyms", and "Hyponyms" with corresponding word suggestions in Arabic and English.

# SAMER Simplification Corpus

## Corpus Annotation:

- Three professional linguists
- Guidelines:

Original Level 5	وليس جميع الحوادث والأحوال تساوي الدم الإنساني الذي لا يوجد أثمن منه، ولا يجب مضاربة أولئك الشعوب الذين يبادرون إلى شن الغارات وفتك بعضهم بعضا على أقل أرب لا يعتد به، أو أدنى خرافة لا بيت لها في رقعة التمدن؛	Not all events and conditions are worth human blood, for there is nothing more valuable than that blood. Nor should one emulate those people who take to launching raids and slaughtering each other for the triflest most dismissible of wants, or for the weakest of superstitions that has no refuge in civilization's domain.
---------------------	--	---

# SAMER Simplification Corpus

## Corpus Annotation:

- Three professional linguists
- Guidelines:
  - 1) Check the readability level assigned to each word
  - 2) Correct any readability leveling errors

Original Level 5	وليس جميع الحوادث والأحوال تساوي الدم الإنساني الذي لا يوجد أثمن منه، ولا يجب <b>مضارعة</b> <sup>1</sup> أولئك الشعوب الذين يبادرون إلى <b>شن</b> <sup>2</sup> الغارات وقتك بعضهم بعضا على أقل <b>أرب</b> <sup>3</sup> لا <b>يعتد</b> <sup>4</sup> به، أو أدنى خرافة لا بيت لها في رقعة التمدن؛	Not all events and conditions are worth human blood, for there is nothing more valuable than that blood. Nor should one <b>emulate</b> <sup>1</sup> those people who take to <b>launching</b> <sup>2</sup> raids and slaughtering each other for the triflest most <b>dismissible</b> <sup>4</sup> of <b>wants</b> <sup>3</sup> , or for the weakest of superstitions that has no refuge in civilization's domain.
---------------------	--	--

# SAMER Simplification Corpus

## Corpus Annotation:

- Three professional linguists
- Guidelines:
  - 1) Check the readability level assigned to each word
  - 2) Correct any readability leveling errors
  - 3) If the document is of:
    - a) **Level 5**, simplify to **Level 4** then to **Level 3**
    - b) **Level 4**, simplify to **Level 3**
    - c) **Level 3**, no simplification is needed
  - 4) Simplification involves minimal replacements, deletions, and insertions to reduce the readability level of the document, while maintaining overall meaning and grammaticality

Original Level 5	وليس جميع الحوادث والأحوال تساوي الدم الإنساني الذي لا يوجد أثمن منه، ولا يجب مضارعة أولئك الشعوب الذين يبادرون إلى شن الغارات وقتك بعضهم بعضا على أقل أرب لا يعتد به، أو أدنى خرافة لا بيت لها في رقعة التمدن؛	Not all events and conditions are worth human blood, for there is nothing more valuable than that blood. Nor should one emulate those people who take to launching raids and slaughtering each other for the triflest most dismissible of wants, or for the weakest of superstitions that has no refuge in civilization's domain.
Level 4	وليس جميع الحوادث والأحوال تساوي الدم الإنساني الذي لا يوجد أثمن منه، ولا يجب أن تشبه أولئك الشعوب الذين يبادرون إلى بدء الغارات وقتك بعضهم بعضا على أقل هدف لا يؤخذ به، أو أدنى خرافة لا بيت لها في رقعة التمدن؛	Not all events and conditions are worth human blood, for there is nothing more valuable than that blood. Nor should one imitate those people who take to starting raids and slaughtering each other for the triflest most negligible of aims, or for the weakest of superstitions that has no refuge in civilization's domains.
Level 3	وليس جميع الحوادث والأحوال تساوي الدم الإنساني الذي لا يوجد أثمن منه، ولا يجب أن تشبه أولئك الشعوب الذين يسرعون إلى بدء الهجمات وقتل بعضهم بعضا على أقل هدف لا يؤخذ به، أو أدنى خرافة لا بيت لها في منطقة التطور؛	Not all events and conditions are worth human blood, for there is nothing more valuable than that blood. Nor should one imitate those people who rush to starting attacks and killing each other for the triflest most negligible of aims, or for the weakest of superstitions that has no refuge in the range of progress.



# SAMER Simplification Corpus

## Inter-Annotator Agreement:

- Selected ~1300 words from each book (~20K words) to be double annotated
- Qualitative Check:
  - ~**6.8%** word-level mismatches when simplifying the original text to Level 4
  - ~**13.2%** word-level mismatches when simplifying Level 4 text to Level 3
  - Most mismatches are due to different lexical simplification choices
  - Annotation mistakes were ~10% of all mismatches

# SAMER Simplification Corpus

## Corpus Statistics (Word-Level):

- The three versions of the texts are comparable in size in terms of the number of words
- Different distributions of readability levels:

	Original		L4		L3	
<b>L0</b>	2,631	1.7%	5,212	3.3%	5,246	3.3%
<b>L1</b>	83,772	52.6%	90,232	56.5%	95,898	59.9%
<b>L2</b>	23,103	14.5%	26,297	16.5%	30,015	18.7%
<b>L3</b>	22,517	14.1%	24,630	15.4%	28,990	18.1%
<b>L4</b>	14,965	9.4%	13,306	8.3%	0	0.0%
<b>L5</b>	9,463	5.9%	0	0.0%	0	0.0%
<b>L6</b>	2,814	1.8%	0	0.0%	0	0.0%
<b>Total</b>	159,265	100%	159,677	100%	160,149	100%

# SAMER Simplification Corpus

## Corpus Statistics (Word-Level):

- The three versions of the texts are comparable in size in terms of the number of words
- Different distributions of readability levels:
  - 8.8% shift from higher to lower readability when simplifying the original text to level 4 (L4)

	Original		L4		L3	
<b>L0</b>	2,631	1.7%	5,212	3.3%	5,246	3.3%
<b>L1</b>	83,772	52.6%	90,232	56.5%	95,898	59.9%
<b>L2</b>	23,103	14.5%	26,297	16.5%	30,015	18.7%
<b>L3</b>	22,517	14.1%	24,630	15.4%	28,990	18.1%
<b>L4</b>	14,965	9.4%	13,306	8.3%	0	0.0%
<b>L5</b>	9,463	5.9%	0	0.0%	0	0.0%
<b>L6</b>	2,814	1.8%	0	0.0%	0	0.0%
<b>Total</b>	159,265	100%	159,677	100%	160,149	100%

# SAMER Simplification Corpus

## Corpus Statistics (Word-Level):

- The three versions of the texts are comparable in size in terms of the number of words
- Different distributions of readability levels:
  - 8.8% shift from higher to lower readability when simplifying the original text to level 4 (L4)
  - 8.3% shift from higher to lower readability when simplifying level 4 (L4) text to level 3 (L3)

	Original		L4		L3	
L0	2,631	1.7%	5,212	3.3%	5,246	3.3%
L1	83,772	52.6%	90,232	56.5%	95,898	59.9%
L2	23,103	14.5%	26,297	16.5%	30,015	18.7%
L3	22,517	14.1%	24,630	15.4%	28,990	18.1%
L4	14,965	9.4%	13,306	8.3%	0	0.0%
L5	9,463	5.9%	0	0.0%	0	0.0%
L6	2,814	1.8%	0	0.0%	0	0.0%
Total	159,265	100%	159,677	100%	160,149	100%

# SAMER Simplification Corpus

## Corpus Statistics (Word-Level):

- The three versions of the texts are comparable in size in terms of the number of words
- Different distributions of readability levels:
  - 8.8% shift from higher to lower readability when simplifying the original text to level 4 (L4)
  - 8.3% shift from higher to lower readability when simplifying level 4 (L4) text to level 3 (L3)
  - 17.1% in the overall readability levels from the original text to level 3 (L3)

	Original		L4		L3	
L0	2,631	1.7%	5,212	3.3%	5,246	3.3%
L1	83,772	52.6%	90,232	56.5%	95,898	59.9%
L2	23,103	14.5%	26,297	16.5%	30,015	18.7%
L3	22,517	14.1%	24,630	15.4%	28,990	18.1%
L4	14,965	9.4%	13,306	8.3%	0	0.0%
L5	9,463	5.9%	0	0.0%	0	0.0%
L6	2,814	1.8%	0	0.0%	0	0.0%
Total	159,265	100%	159,677	100%	160,149	100%

# SAMER Simplification Corpus

## Corpus Statistics (Transformations Statistics):

- Alignment between the original text and its Level 4 simplified version (Original-L4) and Level 4 and Level 3 (L4-L3)
- Edits in terms insertions, deletions, and replacements
- Each edit could have more than one word at a time

	Original-L4		L4-L3	
No Change	152,214	95.5%	145,090	90.8%
1-1	6,430	4.0%	13,508	8.5%
1-m	337	0.2%	476	0.3%
m-1	120	0.1%	235	0.1%
Insert	209	0.1%	354	0.2%
Delete	40	0.0%	122	0.1%

# SAMER Simplification Corpus

## Corpus Statistics (Transformations Statistics):

- Alignment between the original text and its Level 4 simplified version (Original-L4) and Level 4 and Level 3 (L4-L3)
- Edits in terms insertions, deletions, and replacements
- Each edit could have more than one word at a time

	Original-L4		L4-L3	
No Change	152,214	95.5%	145,090	90.8%
1-1	6,430	4.0%	13,508	8.5%
1-m	337	0.2%	476	0.3%
m-1	120	0.1%	235	0.1%
Insert	209	0.1%	354	0.2%
Delete	40	0.0%	122	0.1%

# SAMER Simplification Corpus

## Corpus Statistics (Transformations Statistics):

- Alignment between the original text and its Level 4 simplified version (Original-L4) and Level 4 and Level 3 (L4-L3)
- Edits in terms insertions, deletions, and replacements
- Each edit could have more than one word at a time

	Original-L4		L4-L3	
No Change	152,214	95.5%	145,090	90.8%
1-1	6,430	4.0%	13,508	8.5%
1-m	337	0.2%	476	0.3%
m-1	120	0.1%	235	0.1%
Insert	209	0.1%	354	0.2%
Delete	40	0.0%	122	0.1%



# SAMER Simplification Corpus

## Corpus Statistics (Fragments Statistics):

- Segmented the data based on punctuation
- Each fragment consisted of ~7.5 words on average

	n	n%	Example	
No Change	8,920	43.3%	Original L4 & L3	ولكنها تعلمت في المدارس الفرنسية أيضا، But she studied in French school as well.
Change in L4 only	2,610	12.7%	Original	ونسَمِعُها تَتَدَب وتَتَوَح كالنكلى. And we hear her <b>lamenting and wailing like a bereaved woman</b> .
			L4 & L3	ونسَمِعُها تبكي وتصرخ كفائدة ابنها. And we hear her <b>crying and screaming like one who lost her son</b> .
Change in L3 only	6,369	30.9%	Original L4	يجب أن يترأس الجلسة، He must <b>preside over</b> the session,
			L3	يجب أن يقود الجلسة، He must <b>lead</b> the session,
Change in L4 & L3	2,704	13.1%	Original	أحدهما مسرج ملجم، One of them is saddled and <b>bridled</b> <sub>1</sub> ,
			L4	أحدهما مسرج مربوط، One of them is <b>saddled</b> <sub>2</sub> and <b>tied</b> <sub>1</sub> ,
			L3	أحدهما معد مربوط، One of them is <b>readied</b> <sub>2</sub> and tied,
Total	20,603	100.0%		

# SAMER Simplification Corpus

## Corpus Statistics (Fragments Statistics):

- Segmented the data based on punctuation
- Each fragment consisted of ~7.5 words on average

	n	n%	Example	
No Change	8,920	43.3%	Original L4 & L3	ولكنها تعلمت في المدارس الفرنسية أيضا، But she studied in French school as well.
Change in L4 only	2,610	12.7%	Original	ونسَمِعُها تَتَدَب وتَتَوَح كالنكلى. And we hear her <b>lamenting and wailing like a bereaved woman</b> .
			L4 & L3	ونسَمِعُها تبكي وتصرخ كفائدة ابنها. And we hear her <b>crying and screaming like one who lost her son</b> .
Change in L3 only	6,369	30.9%	Original L4	يجب أن يترأس الجلسة، He must <b>preside over</b> the session,
			L3	يجب أن يقود الجلسة، He must <b>lead</b> the session,
Change in L4 & L3	2,704	13.1%	Original	أحدهما مسرج ملجم، One of them is saddled and <b>bridled</b> <sub>1</sub> ,
			L4	أحدهما مسرج مربوط، One of them is <b>saddled</b> <sub>2</sub> and <b>tied</b> <sub>1</sub> ,
			L3	أحدهما معد مربوط، One of them is <b>readied</b> <sub>2</sub> and tied,
Total	20,603	100.0%		

# Roadmap

- Motivation
- Arabic Linguistic Facts
- The SAMER Project
- The SAMER Simplification Corpus
- **Conclusion**

# Conclusion & Future Work

- Presented the first manually annotated publicly available Arabic corpus for lexical simplification
- Our corpus includes readability level annotations at both the document and word levels
- Two simplified parallels for each text targeting learners are different readability levels
- Future work:
  - Extensions to other genres and domains
  - Models for readability assessment and text simplification

**Thank you!**  
**Q&A**



# **The SAMER Arabic Text Simplification Corpus**

**Bashar Alhafni, Reem Hazim, Juan Piñeros Liberato,  
Muhamed Al Khalil, Nizar Habash**

 [alhafni@nyu.edu](mailto:alhafni@nyu.edu)

 [@balhafni](#)