

جامعة نيويورك أبوظبي



جامعة زايد
ZAYED UNIVERSITY

ZAEBUC: An Annotated Arabic-English Bilingual Writer Corpus

Nizar Habash
New York University Abu Dhabi

David Palfreyman
Zayed University

Introduction

- ZAEBUC: Zayed Arabic-English Bilingual Undergraduate Corpus
 - Zayed University in the United Arab Emirates
 - Undergraduate student essays on set topics
 - Arabic and English texts
 - Bilingual Writer Corpus (not a Parallel Corpus)
 - A combination of Learner Corpus and Genre Corpus
 - Standard Arabic, English
 - Multi-layered annotations: correction, morphology, CEFR
 - Publicly available
- www.zaebuc.org
- The Arabic word زئبق *zi'baq* means 'mercury'

Roadmap

- **Corpus Design & Desiderata**
- **Data Collection**
- **CEFR Annotation**
- **Text Correction**
- **Morphological Annotation**

Corpus Design & Desiderata

- Rich and multilayered annotations
 - Essays written by a cohort of students (to control for variability) in two languages (Arabic and English)
 - Meta-data features: text topic; writer gender; language of schooling; etc.
 - Text corrections, CEFR, and morphological annotations
- High quality annotations
 - Professional annotators, not crowd sourcing
 - Careful inter-annotator checks to control for quality
- Ethical considerations
 - Consent from the writers is required to include the texts
 - Personal information in the texts is redacted
- Wide usability
 - For researchers in education, sociology and sociolinguistics, as well as NLP researchers and developers
- Openness
 - Available publicly for researchers to use and annotate themselves, with minimal restrictions

Data Collection

- IRB Review at Zayed University
- First-year student essays from Fall 2019
- Diagnostic exams in three courses
 - ENG 140: English Composition I
 - ARA 130: Arabic Concepts (the primary composition course)
 - ARA 030: Arabic Preparedness (a zero-credit preparatory course)
- Consent forms and demographic questions
- Prompts

Topic	Prompt
وسائل التواصل الاجتماعي Social Media	وسائل التواصل الاجتماعي وتأثيرها على الفرد والمجتمع. How do social media affect individuals and society?
التسامح Tolerance	كيف نعزز ثقافة التسامح في المجتمع؟ How can the UAE promote a culture of tolerance in society?
التطور الحضاري Development	التطور الحضاري الذي تشهده دولة الإمارات العربية المتحدة What do you think are the most important developments in the UAE at the moment?

Data Collection

- Meta Data

- Anonymous student ID
- School type (government, private, other)
- Language of schooling (Arabic, English, other)
- City of residence
- Gender
- Course (ENG 140, ARA 030, ARA 130)
- Topic
- Date of writing exam
- Length of exam
- Number of days between Arabic and English exams

		Students		Texts	
		397		602	
Gender	Female	353	89%	542	90%
	Male	44	11%	60	10%
High School Type	Government	215	54%	348	58%
	Private	164	41%	229	38%
	Other	18	5%	25	4%
High School Language	English	196	49%	280	47%
	Arabic	183	46%	298	50%
	Other	18	5%	24	4%
Student Language & Topic	Arabic only	9	2%	9	1%
	English only	183	46%	183	30%
	Both	205	52%	410	68%
	<i>Same Topic</i>	149	73%	298	73%
	<i>Diff Topic</i>	56	27%	112	27%
Text Language & Course & Topic	Arabic	214	54%	214	36%
		<i>Social Media</i>		171	80%
		<i>Tolerance</i>		31	14%
		<i>Development</i>		12	6%
	English	388	98%	388	64%
		<i>Social Media</i>		330	85%
		<i>Development</i>		48	12%
		<i>Tolerance</i>		10	3%

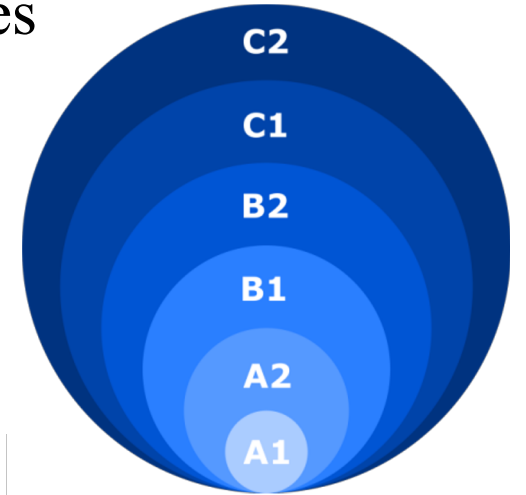
Data Collection

- The vast majority of the students are females.
 - Consistent with the percentage of female students at Zayed University.
- Almost all students contributed to the English sub-corpus
- Two-thirds texts in the corpus are in English
- About half contributed texts in both English and Arabic
- Social Media was the most popular topic by far: 80% in Arabic and 85% in English.

		Students		Texts	
		397		602	
Gender	Female	353	89%	542	90%
	Male	44	11%	60	10%
High School Type	Government	215	54%	348	58%
	Private	164	41%	229	38%
	Other	18	5%	25	4%
High School Language	English	196	49%	280	47%
	Arabic	183	46%	298	50%
	Other	18	5%	24	4%
Student Language & Topic	Arabic only	9	2%	9	1%
	English only	183	46%	183	30%
	Both	205	52%	410	68%
	<i>Same Topic</i>	149	73%	298	73%
	<i>Diff Topic</i>	56	27%	112	27%
Text Language & Course & Topic	Arabic	214	54%	214	36%
	<i>Social Media</i>			171	80%
	<i>Tolerance</i>			31	14%
	<i>Development</i>			12	6%
	English	388	98%	388	64%
	<i>Social Media</i>			330	85%
	<i>Development</i>			48	12%
	<i>Tolerance</i>			10	3%

CEFR Annotation

- The Common European Framework of Reference for Languages
- Six ranked levels from A1 (Beginner), to C2 (Proficient).
- ZAEBUC corpus texts labeled in triplicate
- Examples



	English Example	Arabic Example
C1	Social media is a widely controversial subject with various opinions regarding its negative and positive aspects. While social media has many positive impacts on society, it can also imprint many negative changes on people worldwide. Social media is widely used as a means of communication between people.	في عصرنا الحالي المبني على التكنولوجيا ، تتمتع وسائل التواصل الإجتماعي بأهمية كبيرة، حيث يستصعب على الكثير من الناس العيش من دونها. لدى وسائل التواصل الإجتماعي أثر كبير على حياتنا اليومية وعلينا أن نتفادى الوقوع في سلبيات هذه الآثار. لدى وسائل التواصل الإجتماعي إيجابيات وسلبيات عديدة،
A2	In my opinion I think social media has been the most important thing to everyone. Everyone uses it in the whole part of the earth. It also has a lot of benefits in it, for example knowing about the news and how everything is going on and it is also easier for everyone because ...	قام انتشار الوسائط للتواصل الاجتماعية بشكل كبير وهذا أثر على المجتمع بشكل ايجابي وسلبي من الآثار ايجابية للتواصل الاجتماعي هي التواصل مع الناس بشكل اسهل. ومن ال الآثار ال سلبية هي انتشار الكراهية و الفساد بين الناس.

CEFR Annotation

- Inter-rater Agreement

- The average **pairwise exact** agreement is 47% (Arabic) and 30% (English)
- Arabic Kappa is 0.36 (fair agreement) and English Kappa is 0.16 (slight agreement)
- The average maximum difference in CEFR levels per text is 0.9 (Arabic) and 1.3 (English)
- Average **pairwise fuzzy** agreement (1-level) is 91% (Arabic) and 85% (English)

- CEFR Level Distributions

- The average CEFR is B1 overall
- But Arabic has more B2 than English

- Numerical map

- Overall average is 3.1
- Arabic average is 3.5
- English average is 2.9
- *statistically significant at $p < .001$ using a two-tailed paired T-test on the paired texts by 200 students*

C2=6
C1=5
B2=4
B1=3
A2=2
A1=1

	Level	Arabic	English
Advanced	C1	5%	3%
Upper Intermediate	B2	37%	21%
Intermediate	B1	51%	50%
Pre Intermediate	A2	3%	24%
Beginner	A1	0%	2%
Unassessable		3%	0%

CEFR Annotation

- CEFR Level and Corpus Variables

- Female students vs Male students

- High School Language effect

- High School Type effect

The choice of high school language has a bigger effect on English than on Arabic

		Arabic	English	All
All Students		3.5	2.9	3.1
Gender	Female	3.5	3.0	3.2
	Male	3.4	2.6	2.8
High School Language	Arabic	3.5	2.6	3.0
	English	3.4	3.3	3.3
High School Type	Government	3.5	2.6	3.0
	Private	3.4	3.4	3.4
Topic	Social Media	3.5	3.0	3.2
	Development	3.4	2.5	2.7
	Tolerance	3.5	3.0	3.4

Text Correction

- Spelling and Grammar Correction on all ZAEBUC documents
 - Arabic guidelines (Zaghouani et al., 2014)
 - English guidelines (Dahlmeier et al., 2013)
- Inter-annotator Agreement
 - 26 pairs of texts in English and in Arabic in duplicate
 - Dice Similarity Coefficient
 - Arabic (97.1%) and English (96.7%)
 - The majority of differences are non-erroneous disagreements

Text Correction

- Arabic vs English Errors

Error Type	Count	% of EDI	Raw	Corrected
<i>Punctuation</i>	153	31.2%	important	important;
<i>Typos</i>	79	16.1%	arawnd	around
<i>Verb, Noun Form</i>	64	13.0%	help	helps
<i>Particles</i>	52	10.6%	a bad effect to our	a bad effect on our
<i>Split/Merge/Move</i>	45	9.2%	now a days	nowadays
<i>Capitalization</i>	33	6.7%	social	Social
<i>Pronouns</i>	23	4.7%	you	your
<i>Determiners</i>	19	3.9%	in bad way	in a bad way
<i>Missing/Extra</i>	15	3.1%	there is who	there are some who
<i>Lexical Choice</i>	12	2.4%	sharing a photography	sharing a photograph
<i>Total</i>	495	100.8%		

Error Type	Count	% of EDI	Raw	Corrected
<i>Hamza</i>	140	28.9%	الى	إلى
<i>Wa</i>	139	28.7%	و التواصل	والتواصل
<i>Punctuation</i>	74	15.3%	سيئة	سيئة،
<i>Ta Marbuta</i>	43	8.9%	التعليميه	التعليمية
<i>Typo</i>	37	7.6%	الحظاري	الحضاري
<i>Merge/Split</i>	26	5.4%	من ما	مما
<i>Case</i>	21	4.3%	مكان	مكانا
<i>Diacritical Mark*</i>	14	2.9%	نادراً	نادراً
<i>Gender</i>	13	2.7%	الخاطئي.	الخاطئة.
<i>Word Choice</i>	10	2.1%	اهتمامه في إمارة	اهتمامه بإمارة
<i>Other Features</i>	6	1.2%	يتضررون	يتضرروا
<i>Total</i>	523	108.1%		

Text Correction

- Arabic vs English Errors

	Arabic	English
(a) Text Count	214	388
(b) Raw Word Count	33,376	87,602
Raw Word/Text	156	226
(c) Corrected Word Count	31,661	87,621
Corrected Word/Text	148	226
(d) Exact Match	68.0%	80.3%
Edit	25.7%	17.0%
Delete	6.3%	2.7%
Insert	1.2%	2.7%

Morphological Annotation

- Universal Dependencies (Nivre et al., 2017)
 - Tokenization
 - POS tagging (17 POS tags)
 - Lemmatization
- Automatic Annotation
 - Input is corrected text
 - Arabic: Madamira (Pasha et al., 2014)
 - English: Stanza (Qi et al., 2020)
- Manual Annotation
 - Inter-Annotator Agreement
 - 26 texts in Arabic and English in duplicate
 - English Tokenization (99.98%), POS tagging (99.57%), Lemmatization (99.86%)
 - Arabic Tokenization (99.94%), POS tagging (98.11%), Lemmatization (99.68%)

Morphological Annotation

Raw	Corrected	Edit	WS Tokens	M Tokens	POS	Lemma
the	The	EDIT	The	The	DET	the
social	social		social	social	ADJ	social
media	media		media	media	NOUN	media
didnt	didn't	EDIT	didn't	did+not	AUX+ADV	do+not
affect	affect		affect	affect	VERB	affect
one	one		one	one	NUM	one
country	country		country	country	NOUN	country
or	or		or	or	CCONJ	or
	a	INS	a	a	DET	a
specific	specific		specific	specific	ADJ	specific
group	group		group	group	NOUN	group
of	of		of	of	ADP	of
people,	people;	EDIT	people	people	NOUN	people
			;	;	PUNCT	;
...

Raw	Corrected	Edit	WS Tokens	M Tokens	POS	Lemma
التسامح	التسامح		التسامح	التسامح	NOUN	تسامح
شيء	شيء	EDIT	شيء	شيء	NOUN	شيء
مهم	مهم		مهم	مهم	ADJ	مهم
في	في		في	في	ADP	في
الحياة	الحياة	EDIT	الحياة	الحياة	NOUN	حياة
منه	منه		منه	من+ه	ADP+PRON	من
نتعلم	نتعلم		نتعلم	نتعلم	VERB	تعلم
كيف	كيف		كيف	كيف	ADV	كيف
أن	أن	EDIT	أن	أن	SCONJ	أن
أصبح	أصبح		أصبح	أصبح	VERB	أصبح
أكثر	أكثر	EDIT	أكثر	أكثر	ADJ	أكثر
تعاطف	تعاطفا،	EDIT	تعاطفا	تعاطفا	NOUN	تعاطف
			،	،	PUNCT	،
و، يجب	ويجب	EDIT	ويجب	و+يجب	CCONJ+VERB	وجب
...

Morphological Annotation

- Tokenization

وسیكتبونها
و+ س+ یكتبون +ها
كتب


wsyktbwnhA
w+ s+ yktbwn +hA
katab

and+ will+ they-write +it

	Arabic	English
(a) Text Count	214	388
(b) Raw Word Count	33,376	87,602
Raw Word/Text	156	226
(c) Corrected Word Count	31,661	87,621
Corrected Word/Text	148	226
(d) Exact Match	68.0%	80.3%
Edit	25.7%	17.0%
Delete	6.3%	2.7%
Insert	1.2%	2.7%
(e) WS Token Count	34,235	97,478
WS Token/Text	160	251
(f) Morph Token Count	42,927	98,452
Morph Token/Text	201	254
(g) Al+Morph Token Count	51,609	
Al+Morph Token/Text	241	

Morphological Annotation

- Lexical Similarity




The image displays two word clouds side-by-side, illustrating morphological annotation. The left cloud is in Arabic, and the right cloud is in English. Both clouds consist of various words of different sizes and orientations, representing a collection of terms related to the same theme.

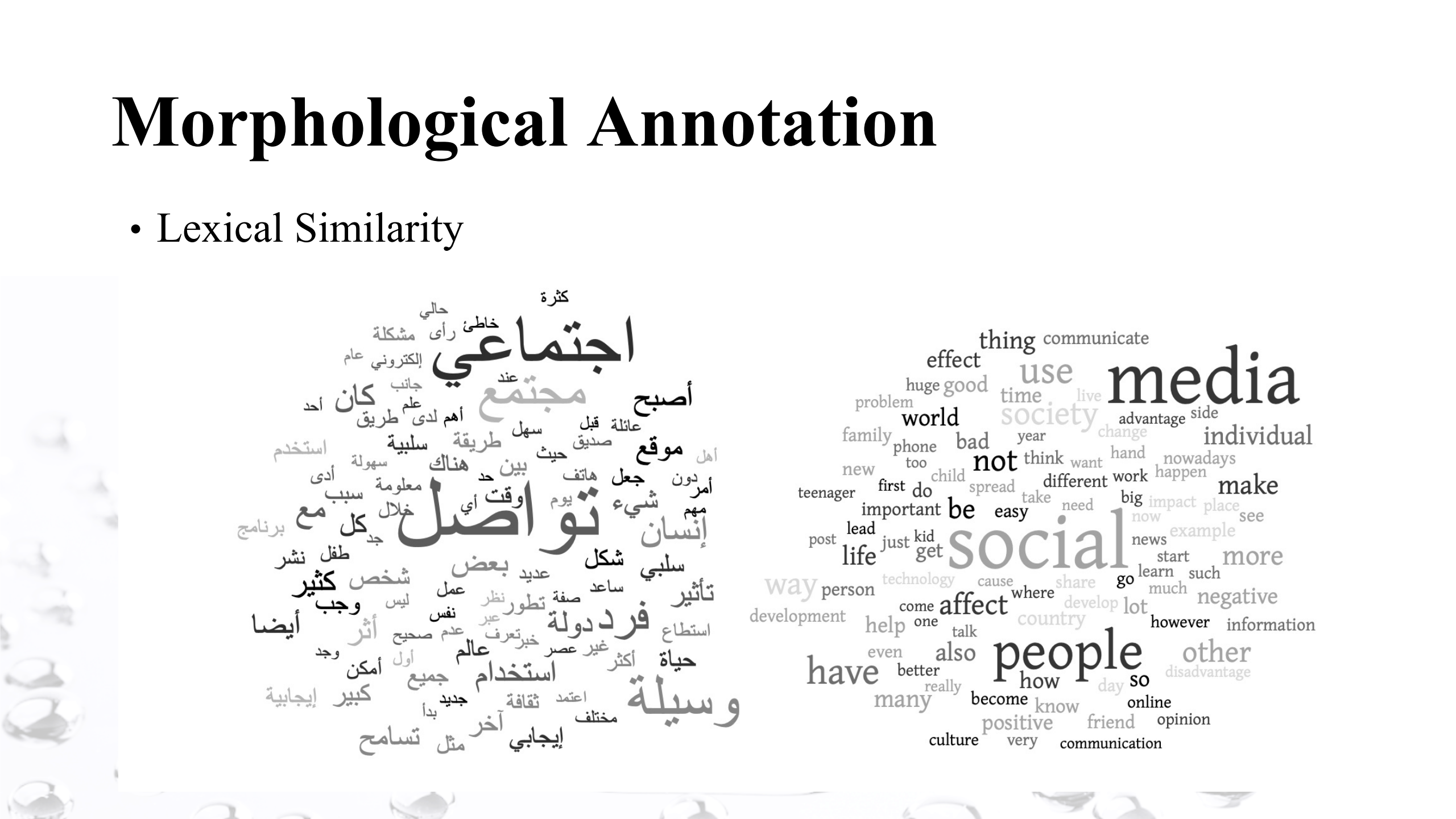
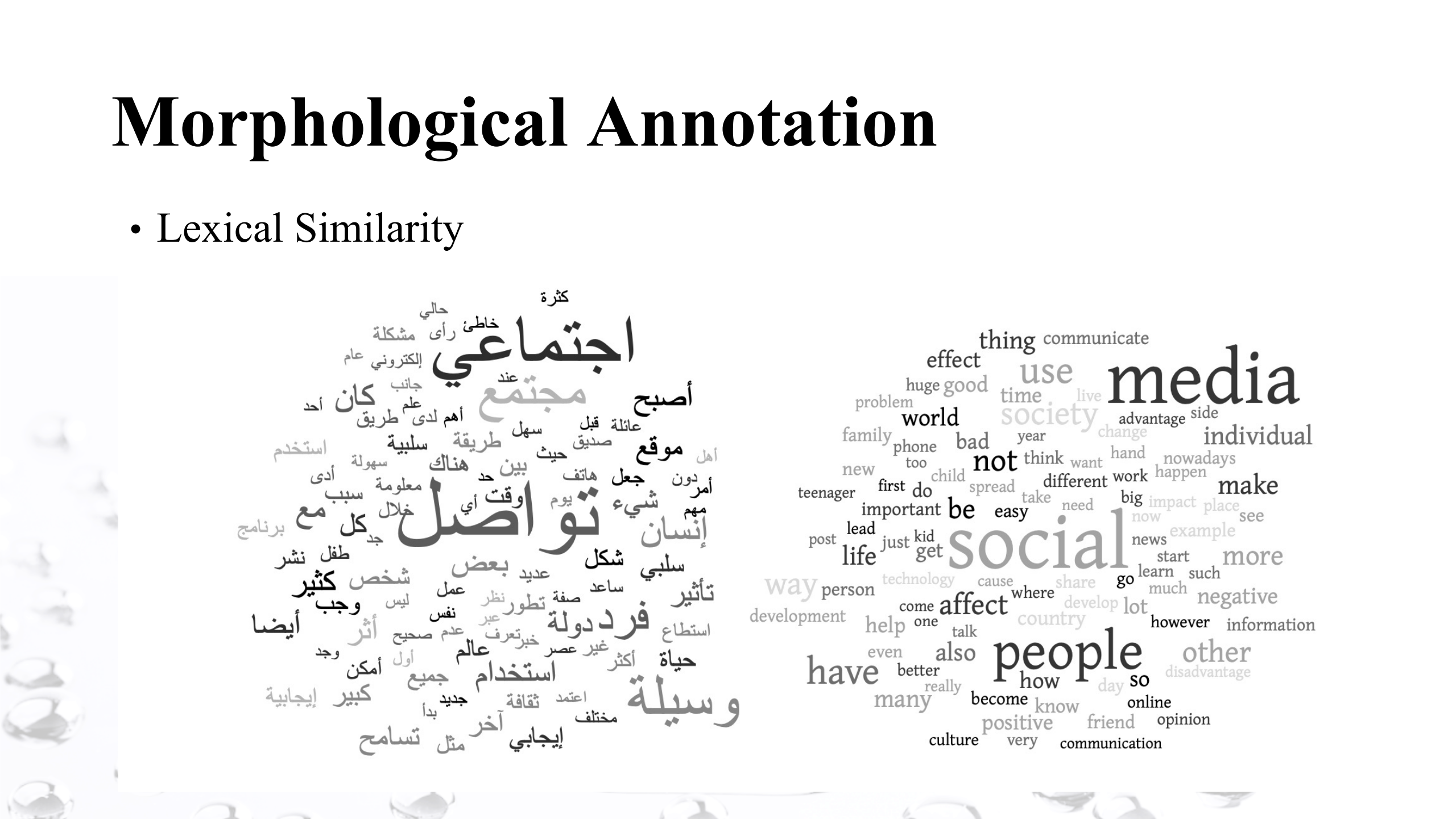
Arabic Word Cloud (Left):

- اجتماعي (Social)
- مجتمع (Community)
- تواصل (Communication)
- وسيلة (Medium)
- أصبح (Became)
- كان (Was)
- أهل (People)
- موقع (Location)
- إنسان (Human)
- شئ (Thing)
- يوم (Day)
- وقت (Time)
- أي (Any)
- كل (All)
- مع (With)
- برنامج (Program)
- شخص (Person)
- كثير (Many)
- أثير (Influence)
- فرد (Individual)
- دولة (Country)
- حياة (Life)
- أكثر (More)
- استخدام (Usage)
- جميع (All)
- أول (First)
- أمكن (Possible)
- كبير (Big)
- إيجابية (Positive)
- جديد (New)
- ثقافة (Culture)
- اعتمد (Depend)
- مختلف (Different)
- إيجابي (Positive)
- آخر (Last)
- مثل (Like)
- تسامح (Tolerance)

English Word Cloud (Right):

- media
- social
- people
- use
- society
- communication
- effect
- thing
- live
- world
- family
- phone
- bad
- not
- think
- want
- hand
- nowadays
- happen
- make
- big
- impact
- place
- see
- example
- news
- start
- more
- learn
- such
- much
- negative
- however
- information
- develop
- lot
- country
- where
- share
- cause
- technology
- person
- way
- development
- help
- one
- talk
- become
- know
- day
- so
- online
- friend
- opinion
- positive
- very
- culture
- communication
- have
- many
- really
- also
- even
- better
- lead
- just
- kid
- get
- life
- post
- important
- do
- first
- child
- spread
- take
- need
- easy
- be
- teenager
- new
- too
- year
- change
- side
- advantage
- individual
- work
- different
- take
- need
- big
- impact
- place
- see
- example
- news
- start
- more
- learn
- such
- much
- negative
- however
- information
- develop
- lot
- country
- where
- share
- cause
- technology
- person
- way
- development
- help
- one
- talk
- become
- know
- day
- so
- online
- friend
- opinion
- positive
- very
- culture
- communication
- have
- many
- really
- also
- even
- better

- # Morphological Annotation
- Lexical Similarity
- 
- The image displays two word clouds side-by-side, illustrating morphological annotation. The left cloud is in Arabic, and the right cloud is in English. Both clouds consist of various words of different sizes and orientations, representing a collection of terms related to the same theme.
- Arabic Word Cloud (Left):**
- اجتماعي (Social)
 - مجتمع (Community)
 - تواصل (Communication)
 - وسيلة (Medium)
 - أصبح (Became)
 - كان (Was)
 - أهل (People)
 - موقع (Location)
 - إنسان (Human)
 - شئ (Thing)
 - يوم (Day)
 - وقت (Time)
 - أي (Any)
 - كل (All)
 - مع (With)
 - برنامج (Program)
 - شخص (Person)
 - كثير (Many)
 - أثير (Influence)
 - فرد (Individual)
 - دولة (Country)
 - حياة (Life)
 - أكثر (More)
 - استخدام (Usage)
 - جميع (All)
 - أول (First)
 - أمكن (Possible)
 - كبير (Big)
 - إيجابية (Positive)
 - جديد (New)
 - ثقافة (Culture)
 - اعتمد (Depend)
 - مختلف (Different)
 - إيجابي (Positive)
 - آخر (Last)
 - مثل (Like)
 - تسامح (Tolerance)
- English Word Cloud (Right):**
- media
 - social
 - people
 - use
 - society
 - communication
 - effect
 - thing
 - communicate
 - live
 - time
 - world
 - problem
 - family
 - phone
 - bad
 - year
 - change
 - advantage
 - side
 - individual
 - nowadays
 - happen
 - work
 - different
 - take
 - need
 - big
 - impact
 - place
 - see
 - example
 - news
 - start
 - more
 - learn
 - such
 - much
 - negative
 - however
 - information
 - develop
 - lot
 - country
 - where
 - share
 - cause
 - technology
 - person
 - way
 - development
 - help
 - one
 - talk
 - become
 - know
 - day
 - so
 - online
 - friend
 - opinion
 - positive
 - very
 - culture
 - communication
 - have
 - many
 - really
 - also
 - even
 - better
 - lead
 - just
 - kid
 - get
 - life
 - post
 - important
 - do
 - first
 - child
 - spread
 - teenager
 - new
 - too
 - family
 - phone
 - bad
 - year
 - change
 - advantage
 - side
 - individual
 - nowadays
 - happen
 - work
 - different
 - take
 - need
 - big
 - impact
 - place
 - see
 - example
 - news
 - start
 - more
 - learn
 - such
 - much
 - negative
 - however
 - information
 - develop
 - lot
 - country
 - where
 - share
 - cause
 - technology
 - person
 - way
 - development
 - help
 - one
 - talk
 - become
 - know
 - day
 - so
 - online
 - friend
 - opinion
 - positive
 - very
 - culture
 - communication
 - have
 - many
 - really
 - also
 - even
 - better



Conclusion & Outlook

- ZAEBUC: Zayed Arabic-English Bilingual Undergraduate Corpus
 - Undergraduate student essays on limited topics
 - Multi-layered annotations: correction, morphology, CEFR
 - Publicly available
- In the future we plan to extend ZAEBUC
 - Add full syntactic representations
 - *(see LREC paper on Camel Treebank)*
 - Add deeper morphological features such as person, gender, and number
 - Conduct diachronic analysis by collecting essays from the same students at a later stage

جامعة نيويورك أبوظبي



Thank you!

www.zaebuc.org

Nizar.Habash@nyu.edu

David.Palfreyman@zu.ac.ae