



# A Benchmark and Scoring Algorithm for Enriching Arabic Synonyms

**Sana Ghanem**

Birzeit University, Palestine

**Mustafa Jarrar**

Birzeit University, Palestine

**Radi Jarrar**

Birzeit University, Palestine

**Ibrahim Bounhas**

Carthage University, Tunisia

# Why do we need Synonyms?

**The importance of synonyms is growing:**

- Application areas: computational linguistics, information retrieval, question answering, and machine translation.
- Essential parts in thesauri, wordnets (Miller et al., 1990), and linguistic ontologies (Jarrar, 2021).

# Notions of Synonymy

- ❖ **Thesauri:** closely related words.
- ❖ **Wordnets:** based on **substitutionability**: “two expressions are synonymous in a linguistic context  $c$  if the substitution of one for the other in  $c$  does not alter the truth value” (Miller et al., 1990).
- ❖ **Linguistic Ontology:** **equivalence relation** (i.e., reflexive, symmetric, and transitive). Two terms are synonyms *iff* they have the exact same concept (i.e., refer, intentionally, to the same set of instances). Thus,  $T_1 =_{Ci} T_2$ . (Jarrar, 2021)

# Two Contributions

**We treat synonymy as a fuzzy value**

1. Experiment: how much linguist agree on synonymy
2. Algorithm to Extract/Extend/Evaluate synsets

# How much linguists agree on synonyms?

## Experiment to measure this agreement:

- Selected 500 synsets from Arabic WordNet, each extended with a number of candidate synonyms (3K synonyms in total).
- Each candidate synonyms was annotated with a fuzzy value by four different linguists.
- The dataset is also used to train our proposed algorithm (i.e., tune its fuzzy model) for extracting synonyms from dictionaries.

# The Dataset Construction

## Synset Selection

500 synsets selected from Arabic WordNet.

### Selection:

#### Part of speech (POS)

1. 350 noun synsets.
2. 140 verb synsets.
3. 10 adjective synsets.

#### Synset's length

1. Length(2), 142 synsets.
2. Length (4), 207 synsets.
3. Length (6), 151 synsets.

➔ Each of the selected synsets was extended with a number of candidate synsets (using our algorithm) – 3k candidates.

➔ Uploaded to Google Sheets (for annotation)

# The Dataset Construction

## Annotation Setup

- Four Linguists (3 training workshops + quiz).
- Each candidate synonyms was annotated with a score based on the linguist's understanding:

مُخَالَفَة   جَلْف   إِتِّحَاد فِدْرَالِي   confederacy   confederation   federation			
a union of political organizations			
مُخَالَفَة ▼	60	نفس الدلالة، الأسلوب ضعيف ، غير شائعة	
إِتِّتِلَاف ▼	80	نفس الدلالة، الأسلوب صحيح ، شائعة الى حد قليل	
إِتِّتِحَاد ▼	100	نفس الدلالة والأسلوب والشيوخ	
جَامِعَة ▼	60	نفس الدلالة، الأسلوب ضعيف ، غير شائعة	

Score	Meaning
100	Same semantics, style, use
90	Same semantics, style, less used
80	Same semantics, style, rarely used
70	Same semantics, style, not used
60	Close semantics, weak style, uncommon
50	Close semantics, not exact purpose
40	Semantically related
30	Semantically related (somehow)
20	Semantically different
10	Semantically very different
0	Semantically unrelated

# The Dataset Construction

## Scoring Guidelines

Scale from 0 to 100 representing the strength of the synonymy relation.

Score	Meaning	
100	Same semantics, style, use	Same semantics
90	Same semantics, style, less used	
80	Same semantics, style, rarely used	
70	Same semantics, style, not used	
60	Close semantics, weak style, uncommon	Close semantics
50	Close semantics, not exact purpose	
40	Semantically related	Related/different semantics
30	Semantically related (somehow)	
20	Semantically different	
10	Semantically very different	
0	Semantically unrelated	



# Results

## Linguists Agreement Evaluation

To measure the (dis/)agreements between linguists, we computed:

1. The Root Mean Squared Error (RMSE).
2. The Mean Average Error (MAE).

	L1		L2		L3		L4		Avg		Algorithm	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
L1			0.19	0.14	0.19	0.14	0.22	0.16	0.13	0.10	0.35	0.30
L2	0.19	0.14			0.16	0.12	0.20	0.15	0.10	0.11	0.31	0.26
L3	0.19	0.14	0.16	0.12			0.20	0.16	0.11	0.08	0.32	0.26
L4	0.22	0.16	0.20	0.15	0.20	0.15			0.13	0.08	0.39	0.34
Avg	0.13	0.10	0.10	0.08	0.11	0.08	0.13	0.11			<b>0.32</b>	<b>0.27</b>
Algorithm	0.35	0.30	0.31	0.26	0.32	0.26	0.39	0.34	<b>0.32</b>	<b>0.27</b>		

Download Dataset: <https://portal.sina.birzeit.edu/synonyms/>

# The Algorithm

## Extract Synonyms from mono or multiple dictionary

**Input:** a mono or multiple dictionary ( $D$ ), and a synset ( $S$ )

dictionary  $D$  consists of set of synsets,  $S_i \in D$ . Each synset is a tuple  $\langle t_1, \dots, t_n \rangle$  of linguistic terms regardless language it belongs to.

---

### Use

---

**Extend** a synset with more synonyms above a given fuzzy value

street | road | طريق | شارع

Extend

Accuracy: 0.14

Level 2  
 Level 3  
 Level 4

**Evaluate/rank** synonyms in a given synset

street | road | طريق | شارع

Evaluate

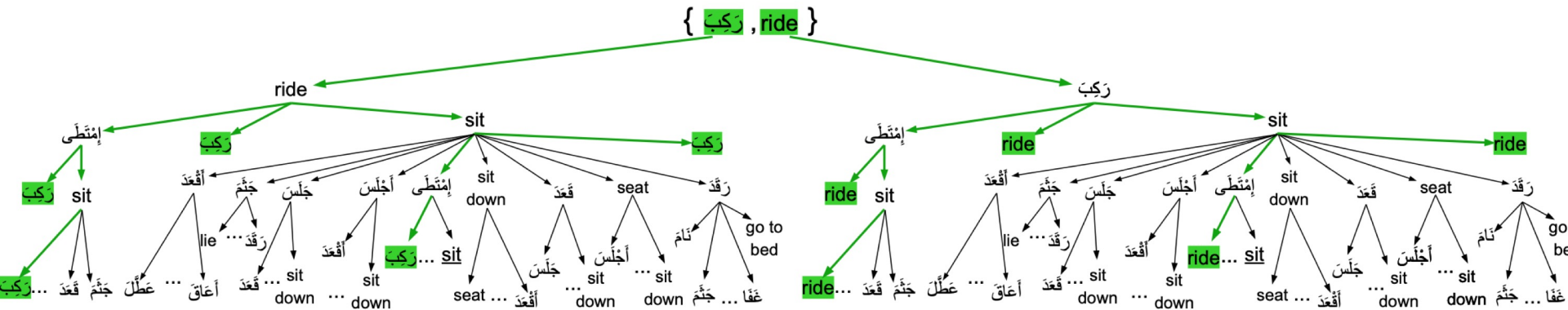


# The Algorithm

## Step 1: Candidate Synonym Extraction

Build directed cyclic graph, from a dictionary: keep expanding until:

- 1) The root node is found, i.e., cycle,
- 2) No more synonyms are found,
- 3) The max  $k$  level is reached.



**Output:** Nodes participating in these cyclic paths are considered candidate synonyms for the given synset.

10 cyclic paths at level 4

- ركب → ride → sit → ركب
- ركب → إمتطى → sit → ركب
- ركب → ride → sit → ركب
- ركب → إمتطى → ركب
- ركب → ride → إمتطى → ركب
- ركب → ride → إمتطى → sit → ركب
- ride → sit → ride
- ride → sit → ركب → ride
- ride → ركب → ride
- ride → إمتطى → ride
- ride → إمتطى → sit → ride

# The Algorithm

## Step 2: Candidate Synonym Selection

$$Fuzzy(f_i) = \emptyset_1 . P_i + \emptyset_2 . Q_i$$

$P_i$  number of cyclic paths that  $c_i$  appears in, divided by the total number of cyclic paths.

$Q_i$  number of root nodes  $t$  that appear in the cyclic paths of  $c_i$ , divided by the total number of terms in the synset  $S$ .

**Output:** a set of candidate synonyms ( $\mathcal{C}$ ), each synonym ( $c_i$ ) is assigned a fuzzy value ( $f_i$ ).

Example:  $Fuzzy(sit) = \frac{6}{10} \times 0.5 + \frac{2}{2} \times 0.5 = 0.8$

$$Fuzzy(إِمْتِظَى) = \frac{6}{10} \times 0.5 + \frac{2}{2} \times 0.5 = 0.8$$

$$\{ sit_{80\%}, إِمْتِظَى_{80\%} \}$$

# The Algorithm

## Parameter Tuning

Our proposed Fuzzy function depends on two constant weights ( $\theta_1$  and  $\theta_2$ ).

We used our Annotated dataset with Arabic WordNet to generate a model with the best results.

we varied the values of the parameters  $\theta_1$  and  $\theta_2$  by selecting their values within the range of [0.1, 0.9] with a step of 0.1 for each parameter.

$\theta_1, \theta_2$		Level 4	Level 3
[0.1, 0.9]	RMSE	0.459	0.377
	MAE	0.375	0.319
[0.2, 0.8]	RMSE	0.408	0.362
	MAE	0.330	0.304
[0.3, 0.7]	RMSE	0.366	0.352
	MAE	0.299	0.296
[0.4, 0.6]	RMSE	0.336	<b>0.349</b>
	MAE	0.280	<b>0.293</b>
[0.5, 0.5]	RMSE	<b>0.321</b>	0.352
	MAE	<b>0.271</b>	0.296
[0.6, 0.4]	RMSE	0.323	0.363
	MAE	0.271	0.304
[0.7, 0.3]	RMSE	0.343	0.382
	MAE	0.272	0.316
[0.8, 0.2]	RMSE	0.378	0.407
	MAE	0.302	0.335
[0.9, 0.1]	RMSE	0.425	0.437
	MAE	0.335	0.357

# Algorithm Evaluation

→ Evaluation of synonyms is known to be difficult (Wu et al., 2003).

## Our Evaluation Methodologies:

- Compare the results with linguists' scores.
  - Behavior (statistically significant)
  - close to the linguists' scores
- Accuracy of the algorithm.

# Algorithm Evaluation

## Testing the algorithm's behavior

**Goal:** whether the scores of the algorithm are statistically significant.

**Statistical test:** One-way ANOVA test (at  $p < 0.05$ ) between the algorithm and the other linguists.

**Prerequisites:** Normality test.

**Result:** The algorithm's scores are not normally distributed.

**Check** the univariate and multivariate outlier analysis.

**Results:** no outliers, *(thus the non-normality of the algorithm's scores are due to skewness in the data and not because of outliers.)*

\* Therefore, **the one-way ANOVA test can be applied.**

**Result of the one-way ANOVA test:** The algorithm has shown to be **not statistically different with the other linguists and their average.**

This test confirms that **the algorithm behaves as a linguist.**



# Algorithm Evaluation

## Comparing the Algorithm with the Baseline

**Goal:** compares the results of our algorithm with the average of the linguists' scores (as a baseline).

**Evaluation metrics:** Root Mean Squared Error (RMSE) and Mean Average Error (MAE).

### Result:

algorithm's scores are **close to the linguists' scores**, which is a good indication that the algorithm scores are realistic.

	L1		L2		L3		L4		Avg		Algorithm	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
L1			0.19	0.14	0.19	0.14	0.22	0.16	0.13	0.10	0.35	0.30
L2	0.19	0.14			0.16	0.12	0.20	0.15	0.10	0.11	0.31	0.26
L3	0.19	0.14	0.16	0.12			0.20	0.16	0.11	0.08	0.32	0.26
L4	0.22	0.16	0.20	0.15	0.20	0.15			0.13	0.08	0.39	0.34
Avg	0.13	0.10	0.10	0.08	0.11	0.08	0.13	0.11			<b>0.32</b>	<b>0.27</b>
Algorithm	0.35	0.30	0.31	0.26	0.32	0.26	0.39	0.34	<b>0.32</b>	<b>0.27</b>		

# Algorithm Evaluation

## Accuracy Evaluation

**Input:** 10K synsets from the Arabic WordNet (AWN).

**Remarks:** (i) no language-specific treatment, (ii) AWN is polysemous

### Prerequisites

1. Calculate the frequency of each synonym (Arabic and English) of all synsets.
2. Select the synonyms with (highest, lowest, average, and random) frequencies in each synset.
3. We considered synsets that contain more than two synonyms, regardless of the language.
4. The terms with the frequency of 1 are not selected.

# Algorithm Evaluation

## Accuracy Evaluation

### Experiment setup

Perform four masking experiments with (highest, lowest, average, and random) frequencies of each synset, at level 3 and level 4.

For each experiment:

1. Masking a synonym with the required frequency in a given synset.
2. Apply the algorithm individually on the masked synset.
3. Measure the accuracy of the algorithm in terms of retrieved words for each synset.

# Algorithm Evaluation

## Accuracy Evaluation

$$\text{Accuracy} = \frac{\text{Top rank correctly retrieved synonyms}}{\text{Sample size}}$$

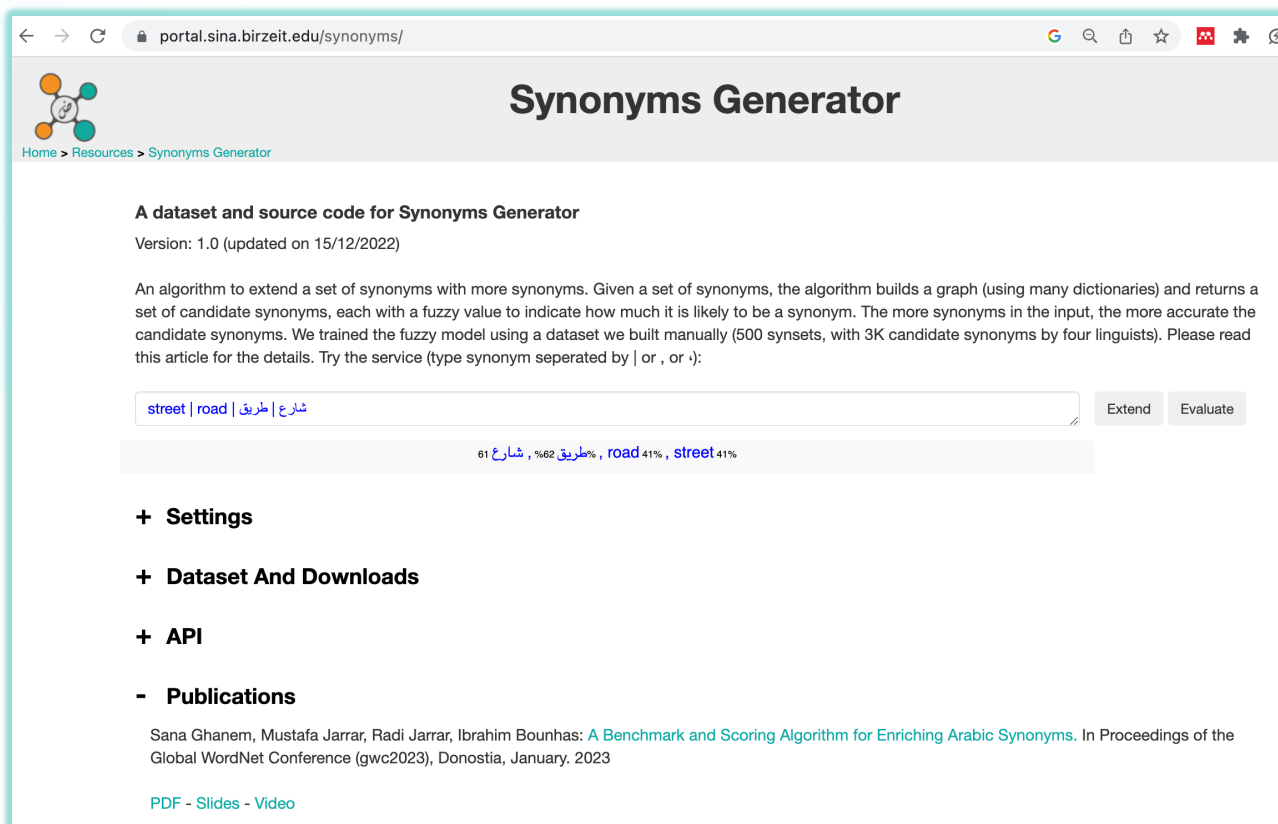
Results:

Experiment	Sample Size	Accuracy at Level 3	Accuracy at Level 4
Exp.1 (Highest)	7, 219	99.1%	95.2%
Exp.2 (Average)	5, 207	98.7%	92.0%
Exp.3 (Lowest)	1, 085	88.4%	62.0%
Exp.4 (Random)	4, 153	98.1%	89.3%

# Summary

- Benchmark dataset fuzzy values.
- Measuring how much linguists dis/agree on synonymy,
- Algorithm to extend/ evaluate synsets (for wordnet, BERT output, etc.)
- No language-specific treatments.

# DEMO



The screenshot shows a web browser window with the URL `portal.sina.birzeit.edu/synonyms/`. The page title is "Synonyms Generator". A breadcrumb trail shows "Home > Resources > Synonyms Generator". The main heading is "A dataset and source code for Synonyms Generator". Below this, it states "Version: 1.0 (updated on 15/12/2022)". A paragraph describes the algorithm: "An algorithm to extend a set of synonyms with more synonyms. Given a set of synonyms, the algorithm builds a graph (using many dictionaries) and returns a set of candidate synonyms, each with a fuzzy value to indicate how much it is likely to be a synonym. The more synonyms in the input, the more accurate the candidate synonyms. We trained the fuzzy model using a dataset we built manually (500 synsets, with 3K candidate synonyms by four linguists). Please read this article for the details. Try the service (type synonym separated by | or , or -):".

Below the text is a search input field containing "street | road | شارع | طريق". To the right of the input are two buttons: "Extend" and "Evaluate". Below the input field, the results are displayed: "61 شارع , %62 طريق , road 41% , street 41%".

On the left side, there is a sidebar with the following sections:

- + Settings
- + Dataset And Downloads
- + API
- Publications

Under the "Publications" section, there is a citation: "Sana Ghanem, Mustafa Jarrar, Radi Jarrar, Ibrahim Bounhas: [A Benchmark and Scoring Algorithm for Enriching Arabic Synonyms](#). In Proceedings of the Global WordNet Conference (gwc2023), Donostia, January. 2023". Below the citation are links for "PDF - Slides - Video".

<https://portal.sina.birzeit.edu/synonyms/>

# References

- Moustafa Al-Hajj and Mustafa Jarrar. 2021. Arab-glossbert: Fine-tuning bert on context-gloss pairs for wsd. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 40–48, Online. INCOMA Ltd.
- Kareem Darwish, Nizar Habash, Mourad Abbas, Hend Al-Khalifa, Huseein T. Al-Natsheh, Houda Bouamor, Karim Bouzoubaa, Violetta Cavalli-Sforza, Samhaa R. El-Beltagy, Wassim El-Hajj, Mustafa Jarrar, and Hamdy Mubarak. 2021. A panoramic survey of natural language processing in the arab worlds. *Commun. ACM*, 64(4):72–81.
- Karim El Haff, Mustafa Jarrar, Tymaa Hammouda, and Fadi Zaraket. 2022. Curras + baladi: Towards a levantine corpus. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2022)*, Marseille, France.
- Mamoun Abu Helou, Matteo Palmonari, and Mustafa Jarrar. 2016. Effectiveness of automatic translations for cross-lingual ontology mapping. *Journal of Artificial Intelligence Research*, 55(1):165–208.
- Mamoun Abu Helou, Matteo Palmonari, Mustafa Jarrar, and Christiane Fellbaum. 2014. Towards building lexical ontology via cross-language matching. In *Proceedings of the 7th Conference on Global WordNet*, pages 346–354. Global WordNet Association.
- Mustafa Jarrar. 2005. *Towards Methodological Principles for Ontology Engineering*. Ph.D. thesis, Vrije Universiteit Brussel.
- Mustafa Jarrar. 2011. Building a formal arabic ontology (invited paper). In *Proceedings of the Experts Meeting on Arabic Ontologies and Semantic Networks*. ALECSO, Arab League.
- Mustafa Jarrar. 2020. *Digitization of Arabic Lexicons*, pages 214–217. UAE Ministry of Culture and Youth.
- Mustafa Jarrar. 2021. The arabic ontology - an arabic wordnet with ontologically clean content. *Applied Ontology Journal*, 16(1):1–26.
- Mustafa Jarrar and Hamzeh Amayreh. 2019. An arabic-multilingual database with a lexico-graphic search engine. In *The 24th International Conference on Applications of Natural Language to Information Systems (NLDB 2019)*, volume 11608 of LNCS, pages 234–246. Springer.
- Mustafa Jarrar, Hamzeh Amayreh, and John P. McCrae. 2019. Representing arabic lexicons in lemon - a preliminary study. In *The 2nd Conference on Language, Data and Knowledge (LDK 2019)*, volume 2402, pages 29–33. CEUR Workshop Proceedings.
- Mustafa Jarrar, Nizar Habash, Faeq Alrimawi, Diyam Akra, and Nasser Zalmout. 2017. Curras: An annotated corpus for the palestinian arabic dialect. *Journal Language Resources and Evaluation*, 51(3):745–775.
- Mustafa Jarrar, Eman Karajah, Muhammad Khalifa, and Khaled Shaalan. 2021. Extracting synonyms from bilingual dictionaries. In *The 11th International Global Wordnet Conference (GWC2021)*, pages 215–222. Global Wordnet Association.
- Mustafa Jarrar, Fadi Zaraket, Rami Asia, and Hamzeh Amayreh. 2018. Diacritic-based matching of arabic words. *ACM Asian and Low-Resource Language Information Processing*, 18(2):10:1–10:21.
- Mustafa Jarrar, Fadi A Zaraket, Tymaa Hammouda, Daanish Masood Alavi, and Martin Waahlsch. 2022. Lisan: Yemeni, irqi, libyan, and sudanese arabic dialect corpora with morphological annotations.
- Eman Naser-Karajah, Nabil Arman, and Mustafa Jarrar. 2021. Current trends and approaches in synonyms extraction: Potential adaptation to arabic. In *Proceedings of the 2021 International Conference on Information Technology (ICIT)*, pages 428–434, Amman, Jordan. IEEE.