The 11th International Global Wordnet Conference,
South Africa, 2021

Extracting Synonyms from Bilingual Dictionaries

Mustafa Jarrar
Birzeit University

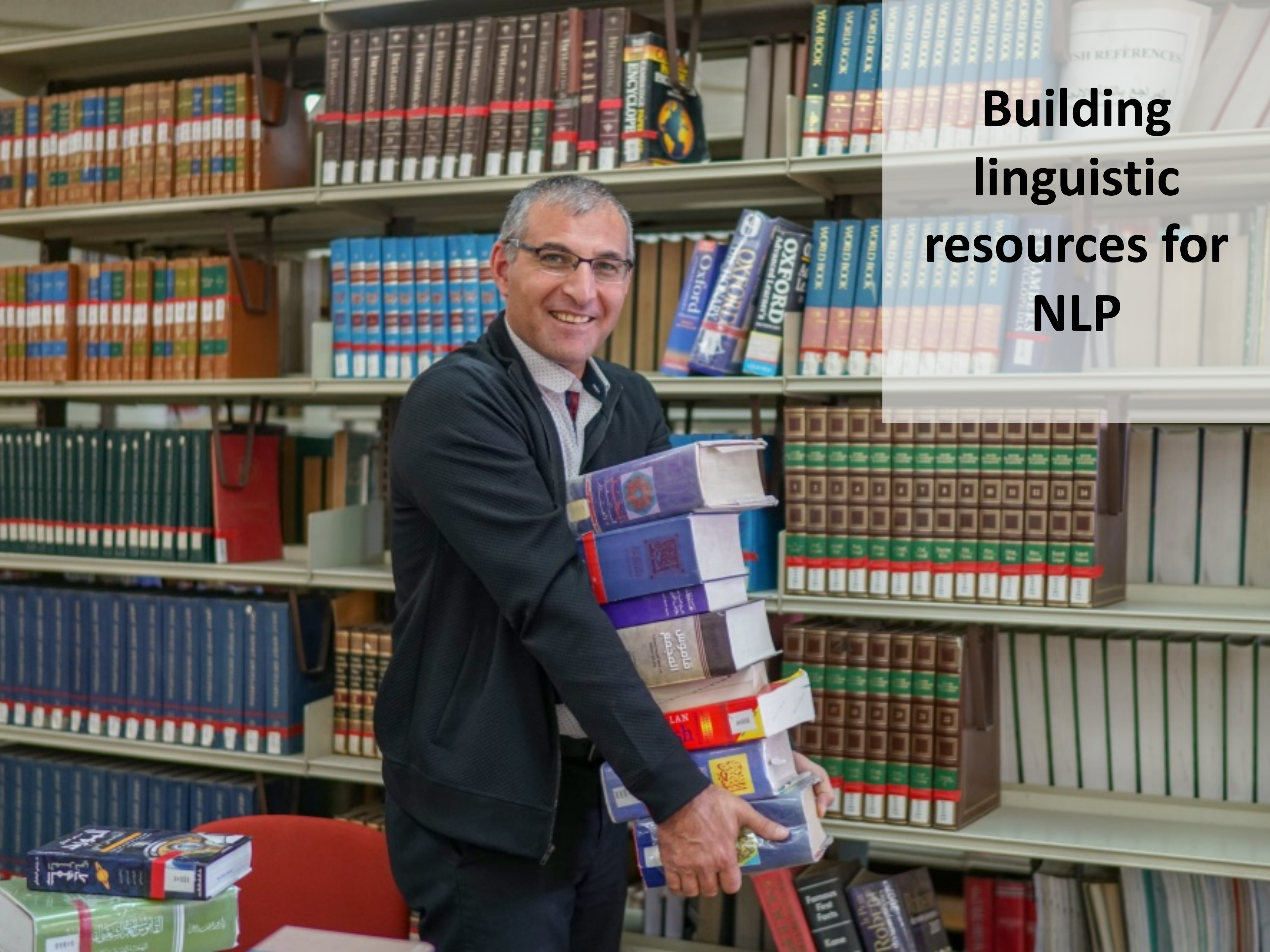
Eman Karajah
Birzeit University

Muhammad Khalifa
Cairo University

Khaled Shaalan
British University in Dubai

جامعة بيرزيت

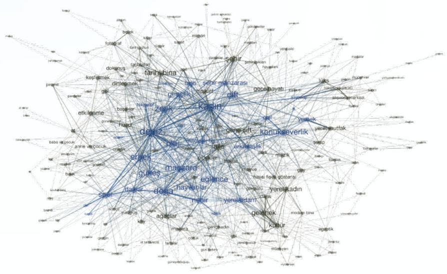
BIRZEIT UNIVERSITY



Building linguistic resources for NLP

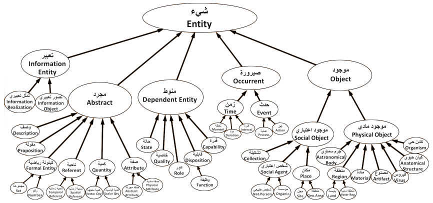
Lexical Resources at Birzeit University

Lexicographic Database



150 lexicons
largest Arabic-multilingual database

Arabic Ontology



Formal Arabic Wordnet
with ontologically clean content

Dialect Corpus

Curras كراس
Corpus for Palestinian Arabic
مدونة اللهجة العامية الفلسطينية

Search [Search]

☑ Stem ☑ MSA ☑ Surface ☑ Gloss
● Whole Word ○ Substring

Annotated dialectal corpus,
each word is annotated with 16 features

Big Linguistic Data Graph

https://ontology.birzeit.edu

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

- **Word embeddings:** words appearing in similar contexts.

- **Thesauri:** closely related words.

- **Wordnets:** based on substitutionablity: “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)

Related Work

Three main tasks related to synonyms extraction:

Wordnet construction

Using other wordnets, machine translation, corpora, emeddings,
(Oliveira and Gomes, 2014), (Ercan and Haziyev, 2019), (Khodak et al., 2017)
(Wu and Zhou, 2003), (Al-Tarouti et al., 2016)

Discovering new translations

Using multilingual translation graphs
(Villegas et al., 2016), (Torregrosa et al., 2019)

Improving existing dictonaries

Analyzing the Ragazzini-Biagi English-Italian dictionary
(Flati and Navigli, 2012)

The Algorithm – Extract Synonyms from bilingual dictionary

- ***Input:*** set of bilingual translation pairs (a_i, e_j)
- ***Do:*** Extract bilingual synonyms, of the form $\{a_1,..,a_k\} = \{e_1,..,e_l\}$.

Step 1: Extract cyclic paths

- Build undirected translation graph, from a dictionary: keep expanding until:
  1) The root node is found, i.e., cycle,
  2) No more translations are found,
  3) The max *k* level is reached.
- *Output:* Nodes participating in the same path are considered candidate synonyms, and converted into bilingual synsets, e.g., $\{a_1, a_2\} = \{e_1, e_2\}$.

Step 2: Consolidation

- Arabic synsets are consolidated (i.e., unioned) if they have the same English synsets
- Similarly, English synsets are consolidated if they have the same Arabic synsets.
- Repeated until no more consolidations are found.

- ***Output:*** the final sets of bilingual synonyms.

Example

Synsets extracted from AWN

أدْغَال | غَاب | غَابَة	forest | wood | woods
غأبَة	forest | timber | timberland | woodland
خَشَب	wood
آلَة نَفْخ	wood | woodwind | woodwind instrument
أَحْرَاش | أدْغَال | حِرْش | دَغْل | غابَة	jungle
نَغْمَة	quality | timber | timbre | tone
نَوْعِيَّة	quality
صَبْغَة | صِبْغَة | دَرَجَة لَوْن | لَوْن خَفِيف	shade | tincture | tint | tone

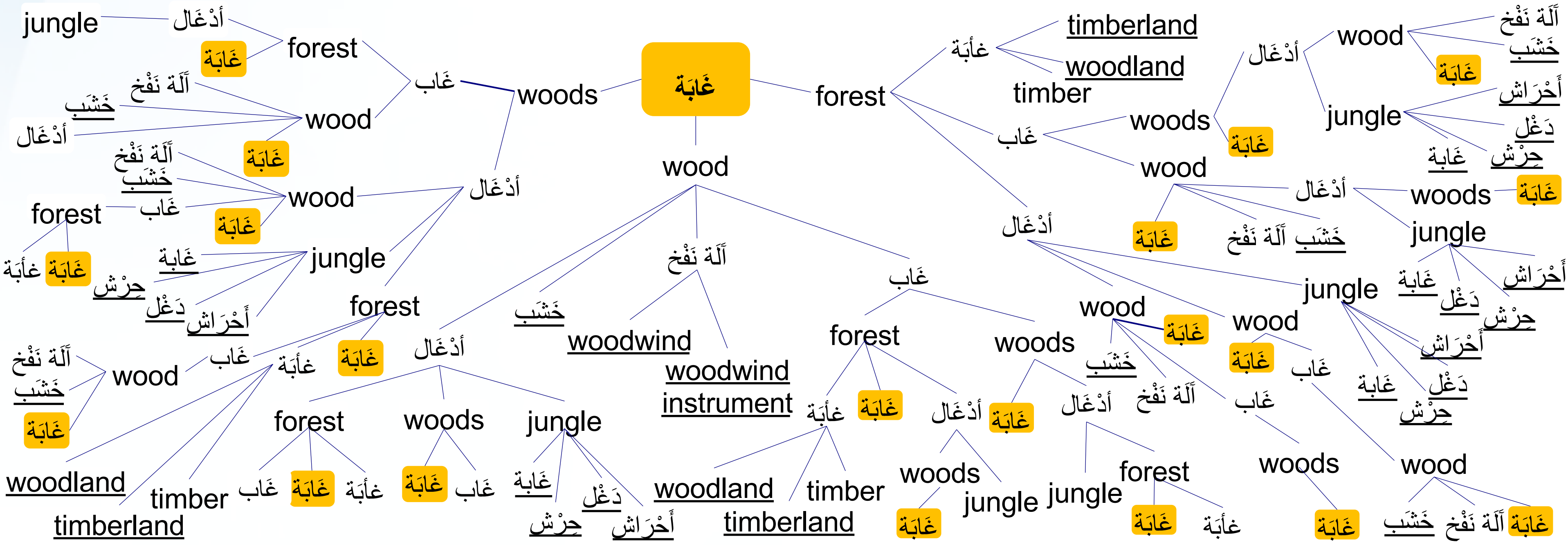
synsets into a flat translation pairs

Bilingual dictionary

آلَة نَفْخ	wood
آلَة نَفْخ	woodwind
آلَة نَفْخ	woodwind instrument
أَحْرَاش	jungle
أدْغَال	forest
أدْغَال	jungle
أدْغَال	wood
أدْغَال	woods
حِرْش	jungle
خَشَب	wood
خَفِيف	shade
خَفِيف	tincture
خَفِيف	tint
خَفِيف	tone
دَرَجَة	shade
دَرَجَة	tincture
دَرَجَة	tint
دَرَجَة	tone
دَغْل	jungle
صَبْغَة	shade
صَبْغَة	tincture
صَبْغَة	tint
صَبْغَة	tone
صِبْغَة	shade
صِبْغَة	tincture
صِبْغَة	tint
صِبْغَة	tone
غأبَة	forest
غأبَة	timber
غأبَة	timberland
غأبَة	woodland
غَاب	forest
غَاب	wood
غَاب	woods
غابَة	jungle
غَابَة	forest
غَابَة	wood
غَابَة	woods
لَوْن	shade
لَوْن	tincture
لَوْن	tint
لَوْن	tone
نَغْمَة	quality
نَغْمَة	timber
نَغْمَة	timbre
نَغْمَة	tone
نَوْعِيَّة	quality

Example (Step 1: Extract cyclic paths)

Translation Graph for ġābaẗ (غَابَة), *k*=7



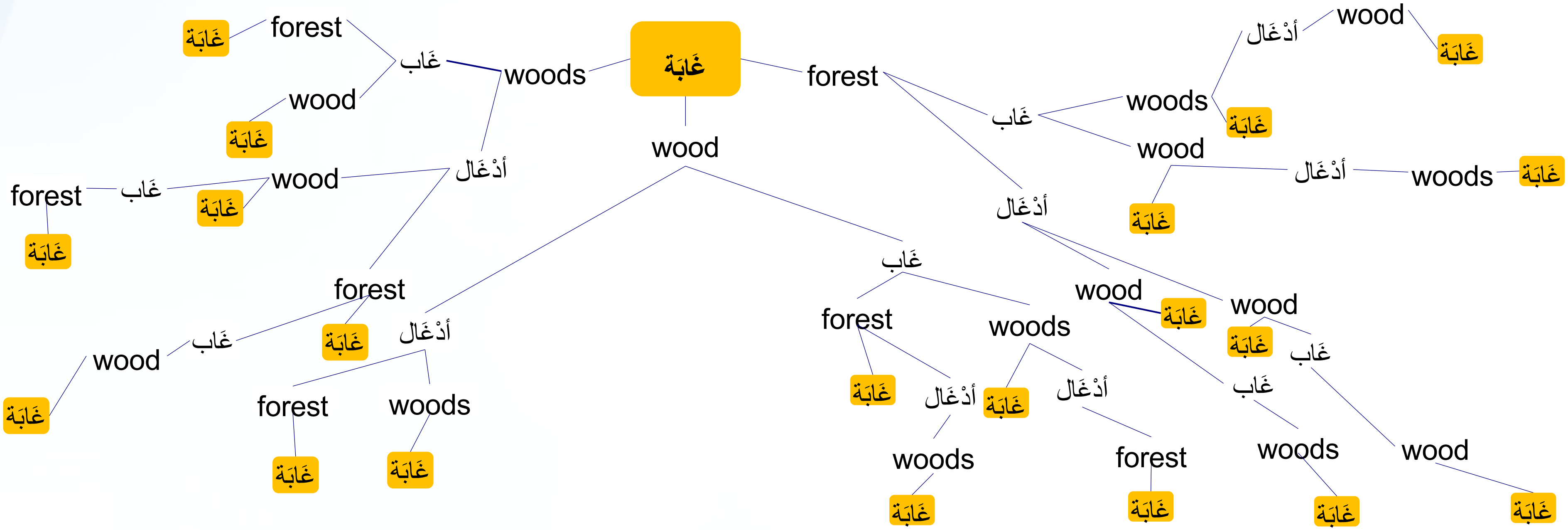
Build undirected translation graph, from a dictionary: keep expanding until:

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

➔ Cyclic paths are Candidate bilingual synsets

Example (Step 1: Extract cyclic paths)

Translation Graph for ġābaẗ (غَابَة), *k*=7



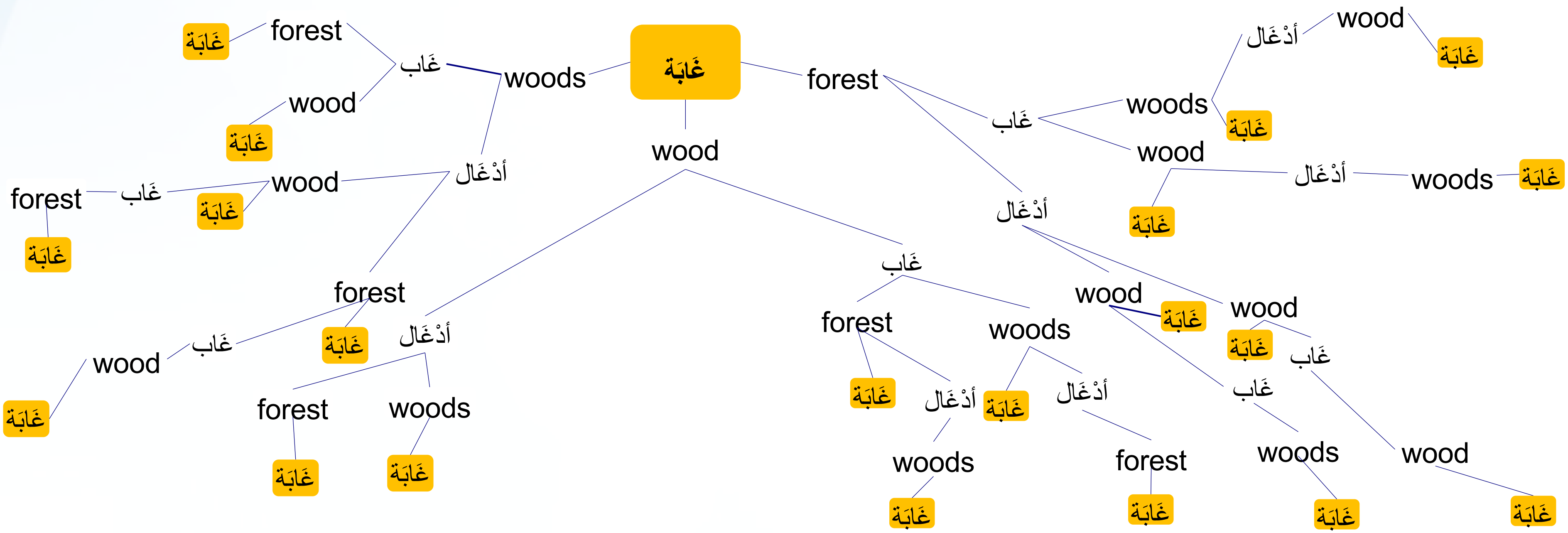
Build undirected translation graph, from a dictionary: keep expanding until:

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

➔ Cyclic paths are Candidate bilingual synsets

Example (Step 1: Extract cyclic paths)

Translation Graph for ġābaẗ (غَابَة), *k*=7



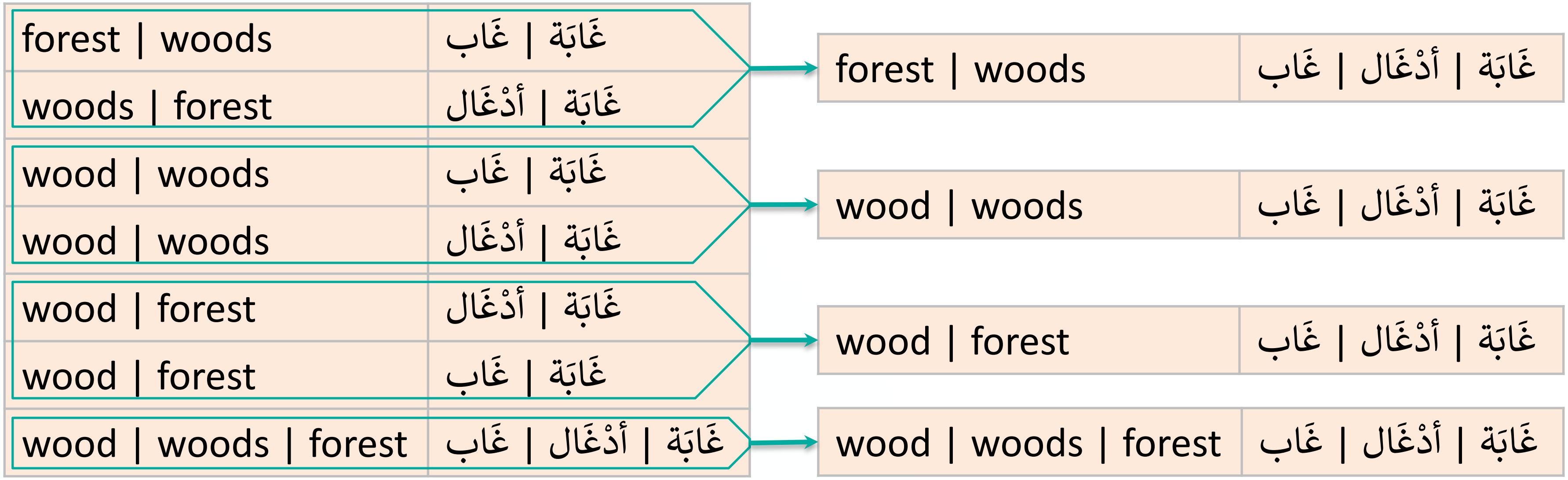
Output: nodes participating in cyclic paths are **candidate bilingual synsets**:

غَاب | غَابَة	Forest | woods
غَاب | أدْغَال | غَابَة	wood | woods | forest
غَاب | غَابَة	wood | woods
أدْغَال | غَابَة	wood | woods
أدْغَال | غَابَة	woods | forest
أدْغَال | غَابَة	wood | forest
غَاب | غَابَة	wood | forest

🡸 duplicates are removed

Example (Step 2: Consolidation)

- Arabic synsets are consolidated (i.e., unioned) if they have the same English synsets
  - Similarly, English synsets are consolidated if they have the same Arabic synsets.
  - Repeated until no more consolidations are found.
  - *Output:* the final sets of bilingual synonyms.



Candidate Synsets

forest | woods	غَابَة | غَاب
woods | forest	غَابَة | أَدْغَال
wood | woods	غَابَة | غَاب
wood | woods	غَابَة | أَدْغَال
wood | forest	غَابَة | أَدْغَال
wood | forest	غَابَة | غَاب
wood | woods | forest	غَابَة | أَدْغَال | غَاب

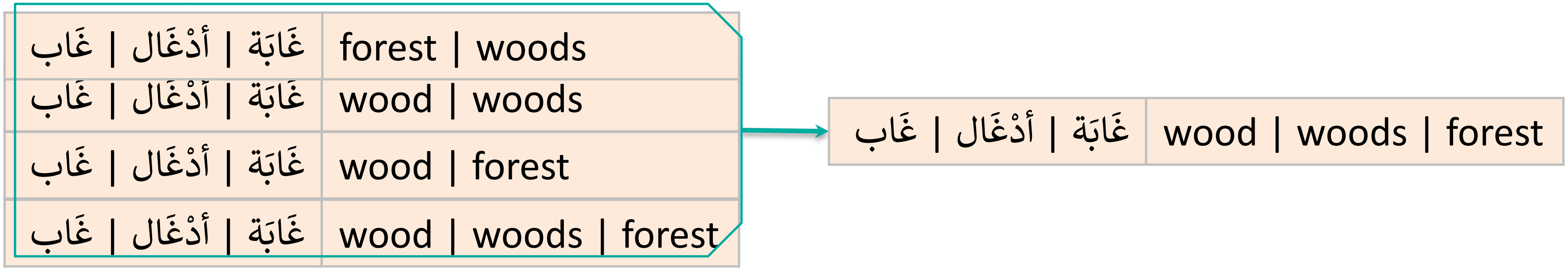
Consolidating Arabic Using English

forest | woods	غَابَة | أَدْغَال | غَاب
wood | woods	غَابَة | أَدْغَال | غَاب
wood | forest	غَابَة | أَدْغَال | غَاب
wood | woods | forest	غَابَة | أَدْغَال | غَاب

Example (Step 2: Consolidation)

- Arabic synsets are consolidated (i.e., unioned) if they have the same English synsets
- Similarly, English synsets are consolidated if they have the same Arabic synsets.
- Repeated until no more consolidations are found.
- *Output:* the final sets of bilingual synonyms.

Consolidating English Using Arabic



غَابَة | أدْغَال | غَاب	forest | woods
غَابَة | أدْغَال | غَاب	wood | woods
غَابَة | أدْغَال | غَاب	wood | forest
غَابَة | أدْغَال | غَاب	wood | woods | forest

غَابَة | أدْغَال | غَاب	wood | woods | forest

no more consolidations needed

Example (Step 2: Consolidation)

- Arabic synsets are consolidated (i.e., unioned) if they have the same English synsets
- Similarly, English synsets are consolidated if they have the same Arabic synsets.
- Repeated until no more consolidations are found.

Output: the final sets of bilingual synonyms.

Final output:

غَابَة | أدْغَال | غَاب	wood | woods | forest

Consolidation Heuristics

The consolidation is designed based on the following **heuristics**:

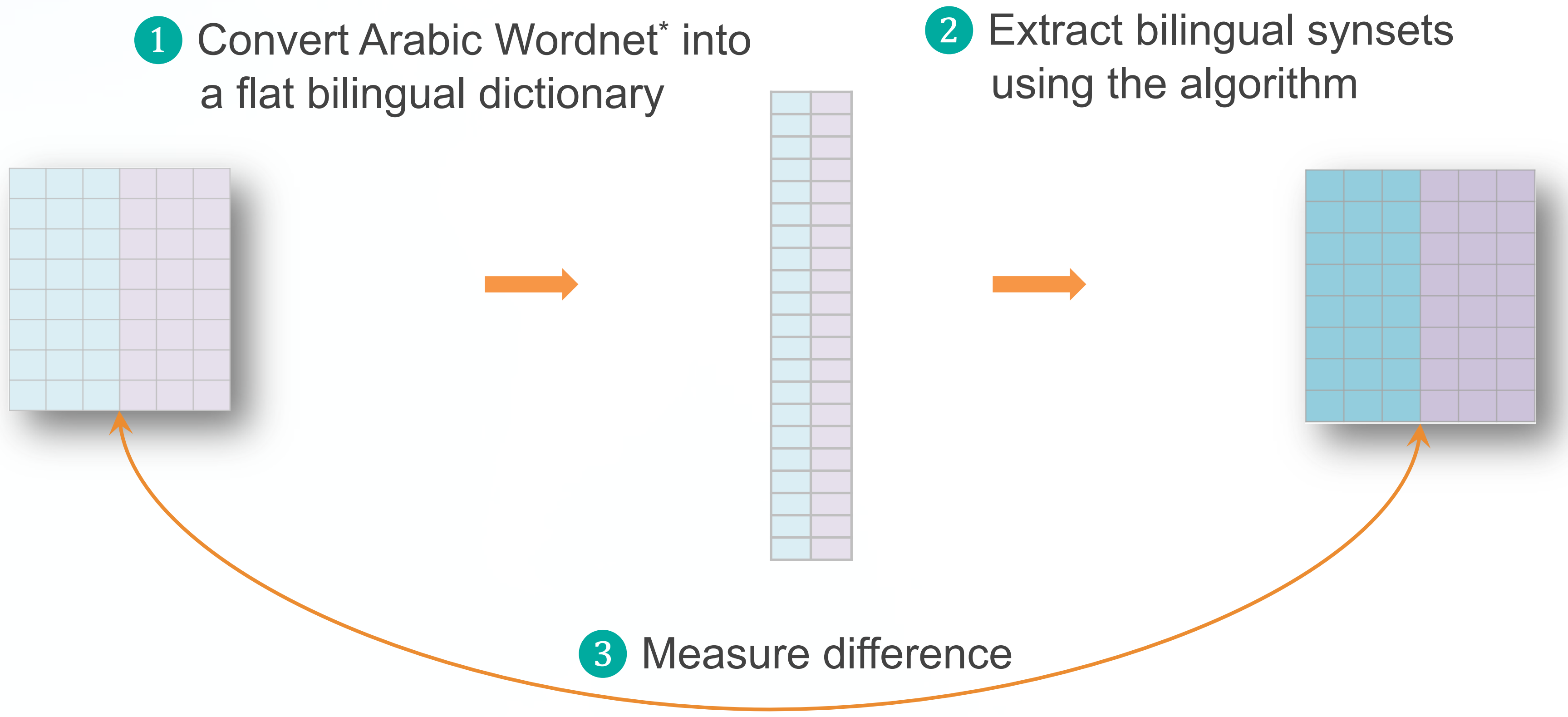
1. It is less likely for bilingual synsets to refer to multiple concepts.

2. It is less likely that a synset is a subset of another synset. Cases like $\{a, b, c, d\}$ and $\{a, b, c\}$ may affect our accuracy.

3. It is less likely for the same English synset, to be translated into multiple Arabic synsets.

Evaluation

- Evaluation of synonyms is known to be difficult (Wu et al., 2003).
- Proposed **Evaluation Methodology**:



1 Convert Arabic Wordnet* into a flat bilingual dictionary

2 Extract bilingual synsets using the algorithm

3 Measure difference

* Arabic Wordnet is very challenging – contains polysemous synsets.

Evaluation

Evaluation matrices:

$$Precision = \frac{\sum_{x \in extracted} max_{y \in AWN} Cosine(x,y)}{|Extracted\ synsets|}$$

$$Recall = \frac{\sum_{y \in AWN} max_{x \in Extracted} Cosine(x,y)}{|AWN|}$$

$$F\text{-}Measure = 2 * \frac{Precision\ .Recall}{Precision + Recall}$$

Results:

	Precision	Recall	F-Measure
k=6, no consolidation	62.5	91.9	74.4
k=6, with consolidation	80.5	84.2	**82.3**
k=8, with consolidation	64.4	84.3	73.0

Remarks: (i) no tuning or any language-specific treatment, (ii) AWN is polysemous

Summary

Conclusions:

- Algorithm to extract synonyms from flat bilingual dictionaries.
- No tuning, No language-specific treatments.
- Good accuracy - although AWN is very polysemous.

Proposed Improvements:

- Fine tuning
- Use part-of-speech, and other morphological features
- Combine words with compatible diacritics or inflections
- Use the algorithm to enrich the Arabic Ontology

Thank You

Mustafa Jarrar
mjarrar@birzeit.edu

→ Email me questions

→ Email me dictionaries to extract you the synonyms

References

Alhafi, D., Deik, D., & Jarrar, M. (2019): Usability Evaluation of Lexicographic e-Services. In Proceedings – 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications, Abu Dhabi (pp.1-7). IEEE. doi:10.1109/AICCSA47632.2019.9035226

Daher, J., & Jarrar, M. (2010). Towards a Methodology for Building Ontologies – Classify by Properties. In Proceedings – 3rd Palestinian International Conference on Computer and Information Technology (PICCIT), Palestine.

Elkateb, S., Black, W., Vossen, P., Farwell, D., Pease A., & Fellbaum, C. (2006). Arabic WordNet and the Challenges of Arabic. In Proceedings – Arabic NLP/MT Conference (pp. 665-670).

Emerson, G. (2020). What are the Goals of Distributional Semantics. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. ACL. (pp. 7436-7453).

Ercan, G., & Haziyev, F. (2019). Synset expansion on translation graph for automatic wordnet construction. Information Processing & Management, 56(1), 130-150.

Helou, M. A., Palmonari, M., & Jarrar, M. (2016). Effectiveness of Automatic Translations for Cross-Lingual Ontology Mapping. Journal of Artificial Intelligence Research, 55, 165-208. doi:10.1613/jair.4789

Helou, M. A., Palmonari, M., & Jarrar, M., Fellbaum, F. (2014). Towards Building Lexical Ontology via Cross-Language Matching. In Proceedings – 7th Conference on Global WordNet. Global WordNet Association. (pp. 346–354). EID: 2-s2.0-84859707947

Jarrar, M., & Meersman, R. (2002). Scalability and Knowledge Reusability in Ontology Modeling. In Proceedings – International Conference on Advances in Infrastructure for Electronic Business, Science, and Education on the Internet (SSGRR 2002s). Scuola Superiore G Reiss Romoli. Rome, Italy.

Jarrar, M. (2005). Towards Methodological Principles for Ontology Engineering. Doctoral dissertation, Vrije Universiteit Brussel, Belgium.

Jarrar, M., (2006). Towards the Notion of Gloss, and the Adoption of Linguistic Resources in Formal Ontology Engineering. In Proceedings – 15th international conference on World Wide Web, (pp.497-503). ACM. doi: 10.1145/1135777.1135850

Jarrar, M. (2011): Building A Formal Arabic Ontology (Invited Paper). In Proceedings – Experts Meeting on Arabic Ontologies and Semantic Networks, Tunis. ALECSO, Arab League.

Jarrar, M., Habash, N., Alrimawi, F., Akra, D., & Zalmout, N. (2016). Curras: An Annotated Corpus for the Palestinian Arabic Dialect. Language Resources and Evaluation, 50(219), 1-31. doi:10.1007/S10579-016-9370-7

Jarrar, M., Zaraket, F., Asia, R., & Amayreh, H. (2018). Diacritic-based Matching of Arabic Words. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 18(2), 1-21. doi: 10.1145/3242177

Jarrar, M., & Amayreh, H. (2019). An Arabic-Multilingual Database with a Lexicographic Search Engine. In Proceedings – 24th International Conference on Applications of Natural Language to Information Systems (NLDB 2019). Lecture Notes in Computer Science (vol. 11608, pp. 234-246). Springer. Doi:10.1007/978-3-030-23281-8_19

Jarrar, M., Amayreh, H., & McCrae, J. (2019): Representing Arabic Lexicons in Lemon – a Preliminary Study. In Proceedings – 2nd Conference on Language, Data and Knowledge, Leipzig, Germany. CEUR-WS (vol. 2402, pp. 29-33).

Jarrar, M. (2021). The Arabic Ontology - An Arabic Wordnet with Ontologically Clean Content. Applied Ontology Journal, IOS Press.

Johnson, D. B. (1975). Finding all the elementary circuits of a directed graph. SIAM Journal on Computing, 4(1), 77-84.

Khodak, M., Risteski, A., Fellbaum, C., & Arora, S. (2017). Automated WordNet construction using word embeddings. In Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications (pp. 12-23).

Lam, K., Tarouti, F., & Kalita J. (2014). Automatically constructing Wordnet synsets. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 106-111).

Miller, J., Beckwith, R., Fellbaum, C., Gross D., & Miller, K. (1990). Introduction to Wordnet: An on-line Lexical Database. International Journal of Lexicography, 3(4), 235-244.

Oliveira, H., & Gomes, P. (2014). ECO and Onto. PT: a flexible approach for creating a Portuguese wordnet automatically. Language resources and evaluation, 48(2), 373-393.

Tarouti, F., & Kalita, J. (2016). Enhancing automatic wordnet construction using word embeddings. In Proceedings – Workshop on Multilingual and Cross-lingual Methods in NLP (pp. 30-34).

Tiziano, F., & Navigli. R. (2012). The CQC algorithm: Cycling in graphs to semantically enrich and enhance a bilingual dictionary. Journal of Artificial Intelligence Research, 43, 135-171.

Torregrosa, D., Mihael, A., Ahmadi, S., & McCrae, J. (2019). TIAD 2019 Shared Task: Leveraging knowledge graphs with neural machine translation for automatic multilingual dictionary generation. Translation Inference Across Dictionaries.

Villegas, M., Melero, M., Gracia J., & Bel, N. (2016). Leveraging RDF graphs for crossing multiple bilingual dictionaries. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) (pp. 868-876).

Wu, H., & Zhow M. (2003). Optimizing synonym extraction using monolingual and bilingual resources. In Proceedings of the second international workshop on Paraphrasing (pp. 72-79).