Natural Language Processing

Lexical Semantics

Mustafa Jarrar

Birzeit University

Watch this lecture and download the slides

Course Page: http://www.jarrar.info/courses/NLP/

More Online Courses at: http://www.jarrar.info

Keywords:
Antonymy, Application Ontologies, Arabic, Arabic NLP, Arabic Ontology, Arabic Wordnet, arabiclanguageday, arabicnlp, arabicontology, Artificial Intelligence, ArtificialIntelligence, Birzeit, Birzeit University, Concept, Corpus, corpus linguistics, Corpus_Tokenizer, datascience, Deep Learning, DeepLearning, diacratics, diacratization, EURO WordNet, Global WordNet, Gloss, Hyponymy, Jarrar, lemmatization, Lemmatizer, lemmatizer, lexical, Lexical Semantics, Lexical_recourse, Linguistic Ontologies, linguistic resources, LinguisticSearchIngine, machinelearning, Meronymy, morphology_tager, Multilingualism, Mustafa Jarrar, Named_Enity, NamedEntityRecognition, Natural Language Processing, NaturalLanguageProcessing, NER, NLP, nlppython, NLProc, NLU, Ontology, parser, Polysemy, POS, pos_tagging, postagging, python, Relation_Extraction, Semantic Relations, Semantic_Relatedness, semantic_tagging, Sina Lab, SinaLab, synonyms, Synonymy, Synset, Thesauri, Toolkit, transliteration, Word_Sense_Disambiguation, Wordnet, WSD, أنطولوجيا لغوية, الانطولوجيا العربية, الترادف اللغوي, التضاد, بيرزيت, تصنيف المعاني, تعدد اللغات, تعدد المعاني, تعريف, جامعة بيرزيت, جرار, حوسبة الدلالة, حوسبة اللغة, سينا, شبكة المفردات, علاقات جزء-كل, مترادفات, مصطفى جرار, معجم ذهني, مفهوم, مكنز, وردنت

Natural Language Processing

Lexical Semantics

In this lecture:

- Part 1: **Linguistic Ontologies vs. Application Ontologies**
- Part 2: What is Lexical Semantics
- Part 3: What is a Concept
- Part 4: Polysemy and Synonymy
- Part 5: Multilingualism
- Part 6: Distributional Semantics

Reading

[J21] Mustafa Jarrar: **The Arabic Ontology - An Arabic Wordnet with Ontologically Clean Content**. Applied Ontology Journal. IOS, 2019.

https://www.jarrar.info/publications/J21.pdf

Application Ontology vs. Linguistic Ontology

The importance of linguistic ontologies is growing rapidly [J21].

Application ontology

To represent the semantics of a certain domain/application, e.g., the Gene Ontology, the FOAF ontology, the Palestinian e-government ontology.

- Each term convey one concept (no polysemy).
- Represents (/Benchmarked to) application's knowledge.
- Used only by a certain application, or a class of applications.

Linguistic ontology

To represent the semantics of terms in a human language, independently of a particular application.

- Each word may convey several concepts (Polysemy).
- Represents common-sense knowledge (/State-of-art scientific discoveries).
- Can be used for general purposes.

Natural Language Processing

Lexical Semantics

In this lecture:

- Part 1: Linguistic Ontologies vs. Application Ontologies
- Part 2: **What is Lexical Semantics**
- Part 3: What is a Concept
- Part 4: Polysemy and Synonymy
- Part 5: Multilingualism
- Part 6: Distributional Semantics

What is Lexical Semantics?

The study of how and what the words of a language denote [J21].

- Whether the meaning of a *lexical unit* is established by looking at its **neighborhood in the semantic net** (by looking at the other words it occurs with in natural sentences), or if the meaning is already locally contained in the lexical unit?
- There are several **theories of the classification and decomposition of word meaning**, the differences and similarities in lexical semantic structure between **different languages**, and **the relationship of word meaning to sentence meaning** and syntax.

Lexical Semantics ➔ focuses on the mapping of words to concepts.

Lexical item: a single word or chain of words that forms the basic elements of a language's lexicon (vocabulary). E.g., "cat", "traffic light", "take care of", "by-the-way“, etc.

What is Lexical Semantics?

- There are **different theories and approaches** in defining the relation between a lexical unit and its meaning(s). For example: can we understand the meaning independently of a sentence? can we understand the meaning independently of the grammar (morphology)? and so on.

- Such theories and approaches are: Prestructuralist semantics, Structuralist and nostructuralist semantics, interpretative semantics and generative semantics, cognitive semantics.

- **In this lecture,** we don't investigate these theories, but rather, we study the "meaning" from a computational and engineering viewpoints, so to enable computer applications ➔ Based on [J21].

Natural Language Processing

Lexical Semantics

In this lecture:

- Part 1: Linguistic Ontologies vs. Application Ontologies
- Part 2: What is Lexical Semantics
- Part 3: **What is a Concept**
- Part 4: Polysemy and Synonymy
- Part 5: Multilingualism
- Part 6: Distributional Semantics

Natural Language Processing

Lexical Semantics

In this lecture:

- Part 1: Linguistic Ontologies vs. Application Ontologies
- Part 2: What is Lexical Semantics
- Part 3: **What is a Concept**
- Part 4: Polysemy and Synonymy
- Part 5: Multilingualism
- Part 6: Distributional Semantics

What is a concept?

ISO TC37 definitions:

تعريف المفهوم حسب مؤسسة المعايير الدولية

3.2.1 concept

unit of knowledge created by a unique combination of **characteristics** (3.2.4)

NOTE Concepts are not necessarily bound to particular languages. They are, however, influenced by the social or cultural background which often leads to different categorizations.

3.2.2 individual concept

concept (3.2.1) which corresponds to only one **object** (3.1.1)

NOTE 1 Examples of individual concepts are 'Saturn', 'the Eiffel Tower'.

NOTE 2 Individual concepts are usually represented by **appellations** (3.4.2)

This ISO definition is based on Eugen Wuster work [W03] who argued: **concepts and objects are both *thoughts* existing in our minds, rather than in reality**.

هذا التعريف منطلق من عمل **إيغن ويستر** الذي يعرف المفهوم والمدلول (الماصدق) أفكار في دماغنا، وليس بالواقع . تم انتقاد هذا التعريف من بعض الفلاسفة، بان هذا غير مفيد لانه لامكن معرفه ما في دماغنا، واقترح باري سمث: استعمال **الكليات** (universal) بدلا من المفاهيم.

This definition was largely criticized [S04, SCT04, S06]: (See [J21])

Constructing concepts as “thoughts” does not help us to benchmark the correctness of our concept system - as we cannot gain access to the interiors of each other’s brains. Smith Suggested to use **Universal** instead of concept.

What is a concept?

ISO TC37 definitions:

تعريف المفهوم حسب مؤسسة المعايير الدولية

3.2.1 concept

unit of knowledge created by a unique combination of **characteristics** (3.2.4)

NOTE Concepts are not necessarily bound to particular languages. They are, however, influenced by the social or cultural background which often leads to different categorizations.

3.2.2 individual concept

concept (3.2.1) which corresponds to only one **object** (3.1.1)

NOTE 1 Examples of individual concepts are 'Saturn', 'the Eiffel Tower'.

NOTE 2 Individual concepts are usually represented by **appellations** (3.4.2)

This ISO definition is based on Eugen Wuster work [W03] who argued: **concepts and objects are both *thoughts* existing in our minds, rather than in reality**.

هذا التعريف منطلق من عمل **إيغن ويستر** الذي يعرف المفهوم والمدلول (الماصدق) أفكار في دماغنا، وليس بالواقع . تم انتقاد هذا التعريف من بعض الفلاسفة، بان هذا غير مفيد لانه لامكن معرفه ما في دماغنا، واقترح باري سمث: استعمال **الكليات** (universal) بدلا من المفاهيم.

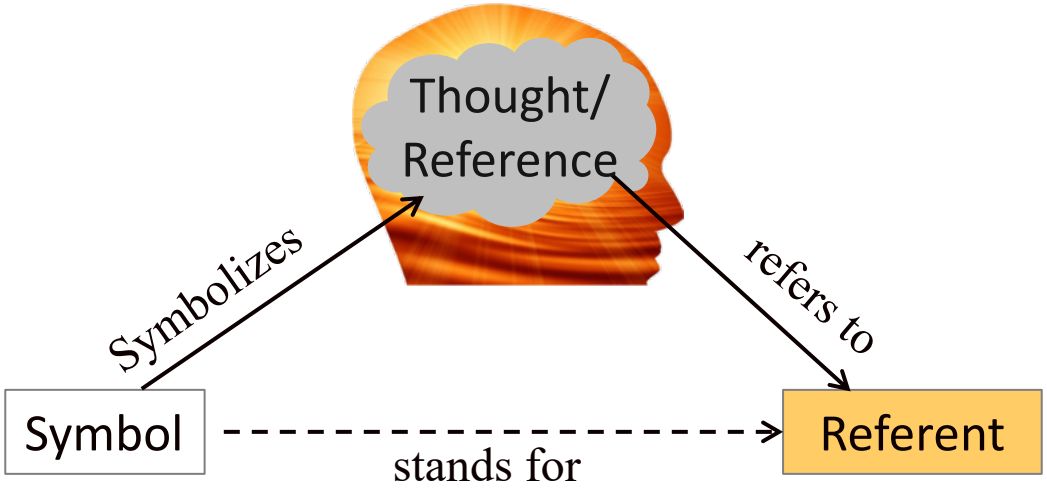
This definition was largely criticized [S04, SCT04, S06]: (See [J21])

Constructing concepts as “thoughts” does not help us to benchmark the correctness of our concept system - as we cannot gain access to the interiors of each other’s brains. Smith Suggested to use **Universal** instead of concept.

The Semiotic Triangle

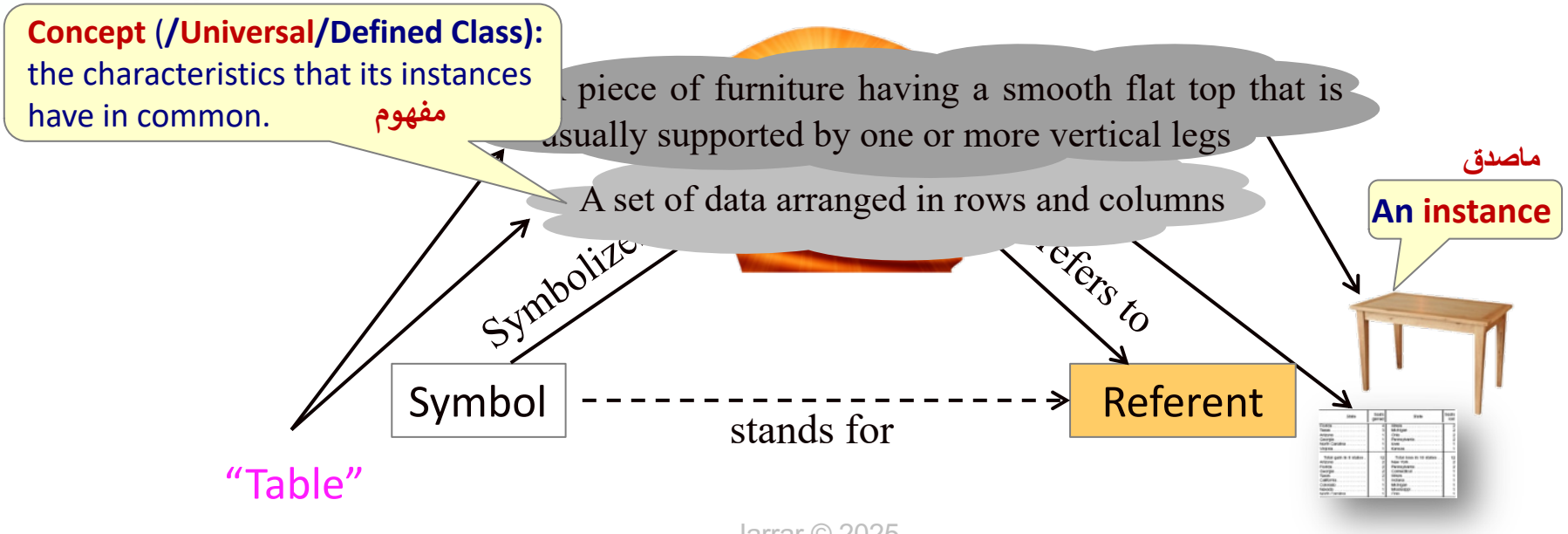
Ogden CK, Richards IA. **The Meaning of Meaning**. 3rd ed. New York, 1930.

- A psychological account (theory of causality)
- A symbolic representation does not refer directly to an object, but indirectly, through a ‘thought or reference’ in mind.



What is a concept

- No concept without instances, and no instance can be a concept at the same time [J11, J05].
- Two concepts with exactly the same extension are the same concept [J11].
- Two terms lexicalizing the same concept (same extension) are synonyms [J05].



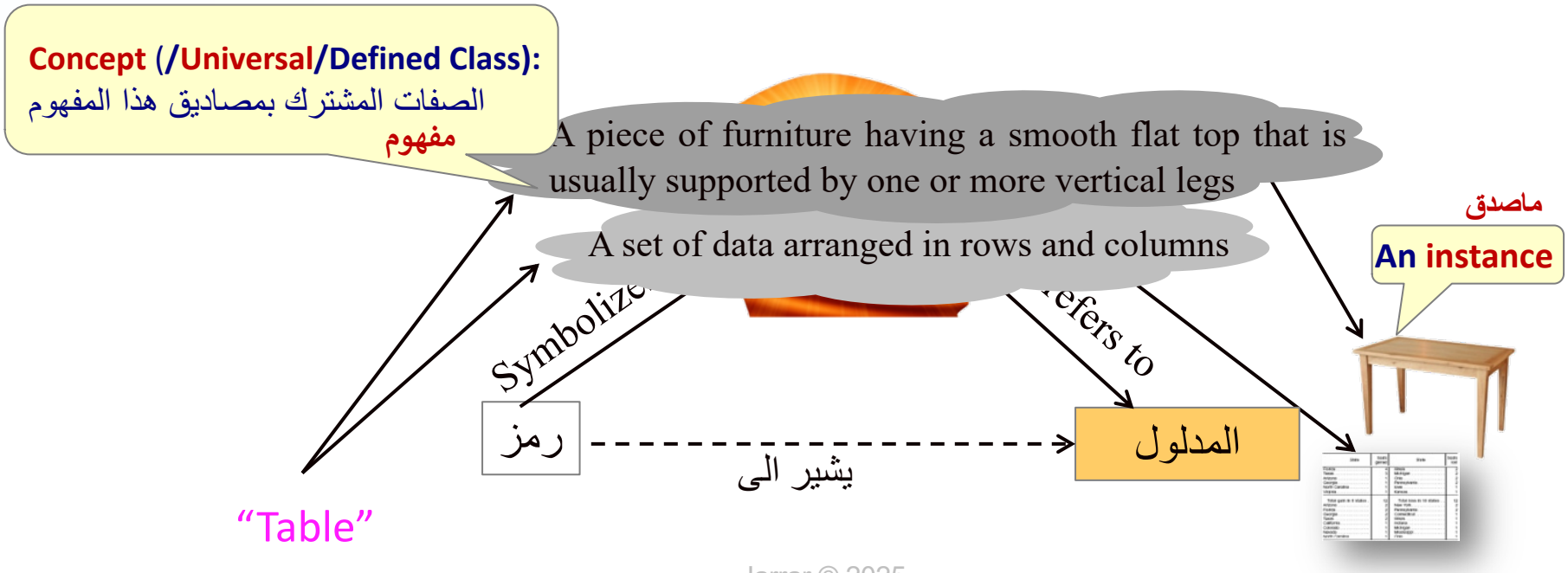
المفهوم والماصدق (Concept vs Instance)

المفهوم هو مجموعة من المصاديق (لها صفات مشتركة).

لا يوجد مفهوم لا يوجد له ماصدق. والماصدق لا يمكن ان يكون مفهوم.

اذا وجد مفهومين لهما نفس المصاديق فهما نفس المفهوم.

اذا وجد كلمتين تشيرا الى نفس المفهوم – أي نفس المصاديق – فهما مترادفتان

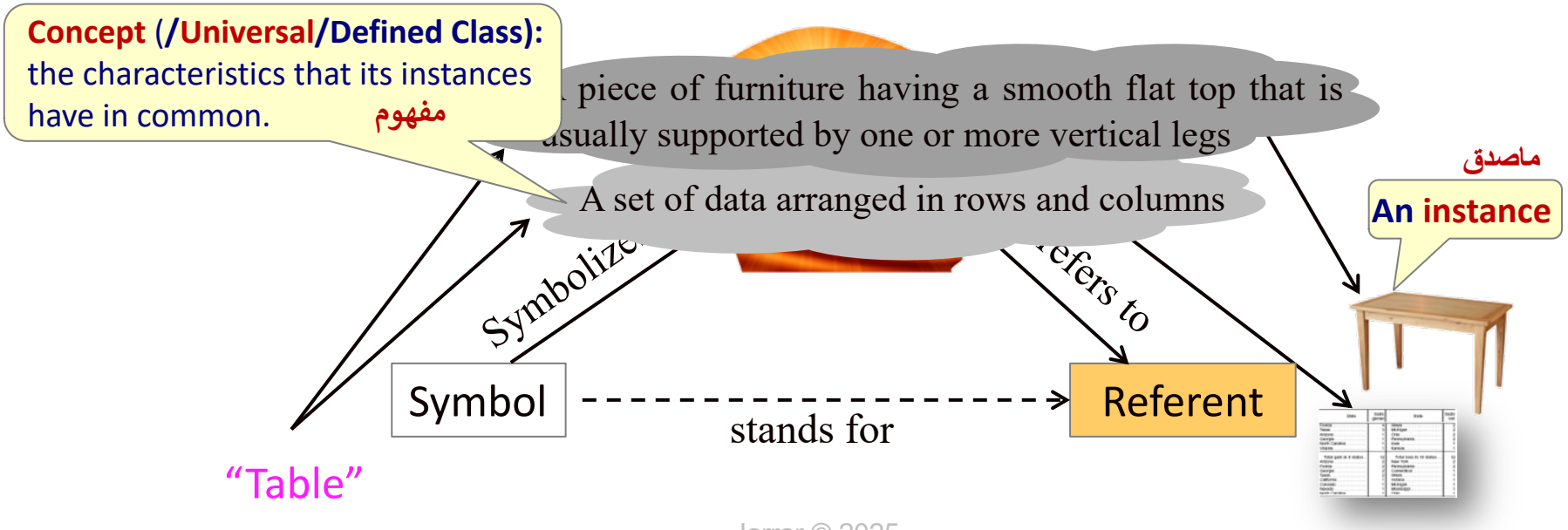


Benchmarking Concepts

How to judge whether the content of the ontology is correct? [J21]

(e.g., Fungus *IsA* Organism? Virus *IsA* Organism? Democracy is political System? …)

- Wusteren/**conceptualist** viewpoint: benchmarked to your *perception*.
- **Realist** Viewpoint: benchmarked to scientific discoveries.



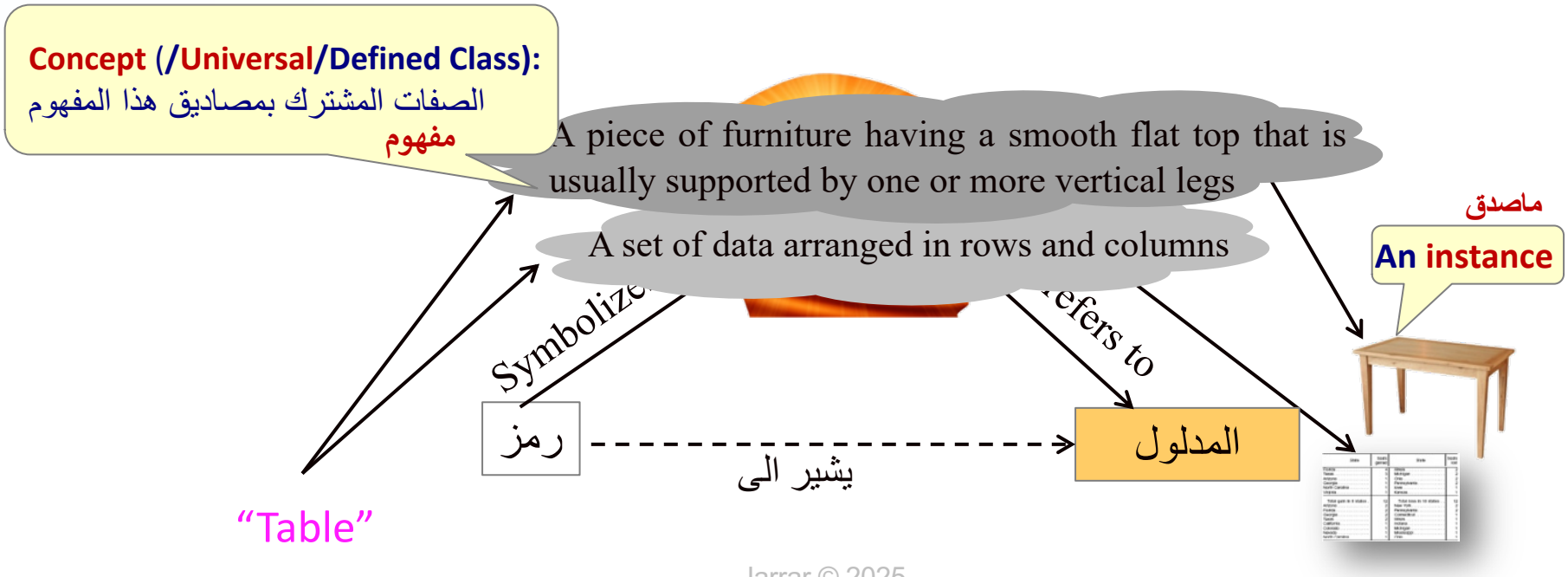
مرجعية العريف المعنى

كيف نحكم ما اذا كان (تعريف المعنى) فعلا صحيح؟ انظر ([J21])

مثلا: الفطر: هو كائن حي ....؟ الفيروس: هو كائن حي: ... ؟ الديمقراطية: نظام حكم ....؟

من وجهة نظر الذهنين (conceptualist) وحسب ويستر: المرجع هو فهمنا (ما في دماغنا)

من وجهة نظر واقعية (Realist) وحسب سمث: المرجع هو ما توصلت اليه العلوم (بالتجارب العلمية)



Number of Terms vs Concepts in a Language

How concepts are named and communicated?

ركوع

قرفصاء

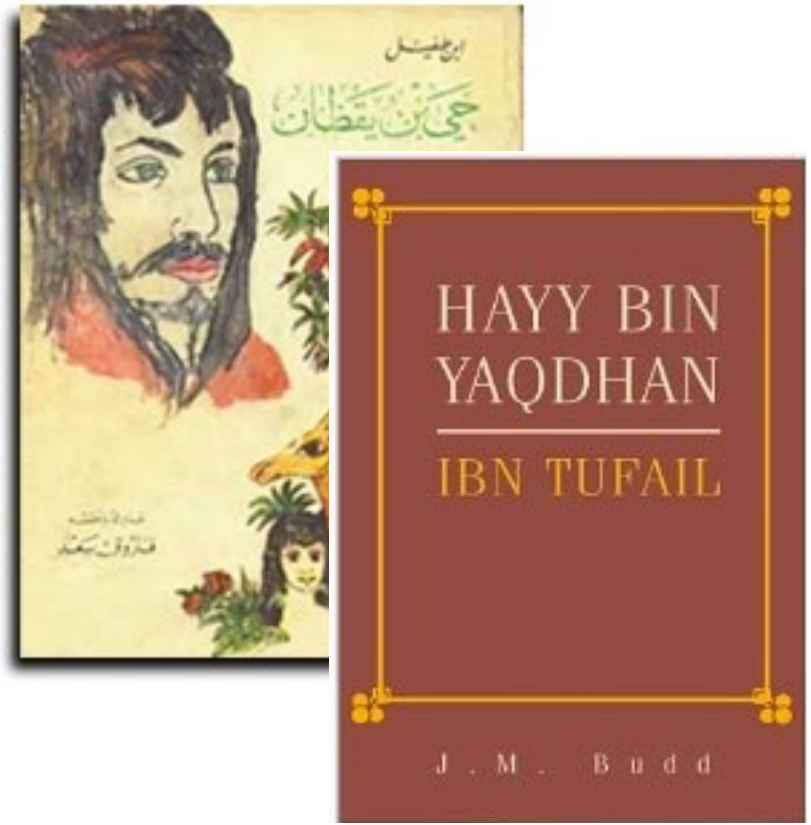
سجود

??

Why there is no word in Arabic to describe this situation?

We give names to the concepts we use more.
Some societies love to give/invite names to concepts (as Arabs in the past)

Recommended Reading



Ḥayy Bin Yaqdhan Novel

Ḥayy ibn Yaqẓān (Arabic: حي ابن يقظان"Alive, son of Awake"; Latin: Philosophus Autodidactus "The Self-Taught Philosopher"; English: The Improvement of Human Reason: Exhibited in the Life of Hai Ebn Yokdhan), the first Arabic novel, was written by Ibn Tufail (also known as Aben Tofail or Ebn Tophail), a Moorish philosopher and physician, in early 12th century Islamic Spain. The novel was itself named after an earlier Arabic allegorical tale and philosophical romance of the same name, written by Avicenna (Ave Cena) in early 11th century,[SO96] though they had different stories.[D92]
-wikipedia

Natural Language Processing

Lexical Semantics

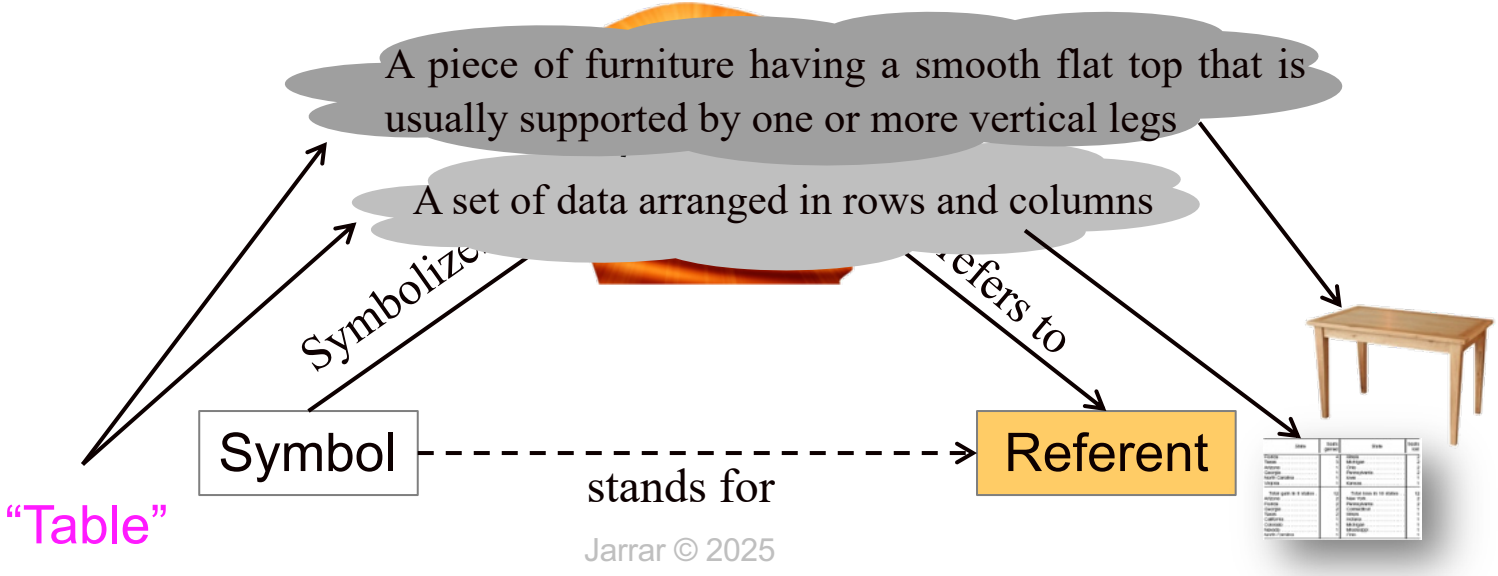
In this lecture:

- ❑ Part 1: Linguistic Ontologies vs. Application Ontologies
- ❑ Part 2: What is Lexical Semantics
- ❑ Part 3: What is a Concept
- ❑ Part 4: **Polysemy and Synonymy**
- ❑ Part 5: Multilingualism
- ❑ Part 6: Distributional Semantics

Polysemy

المشترك اللفظي

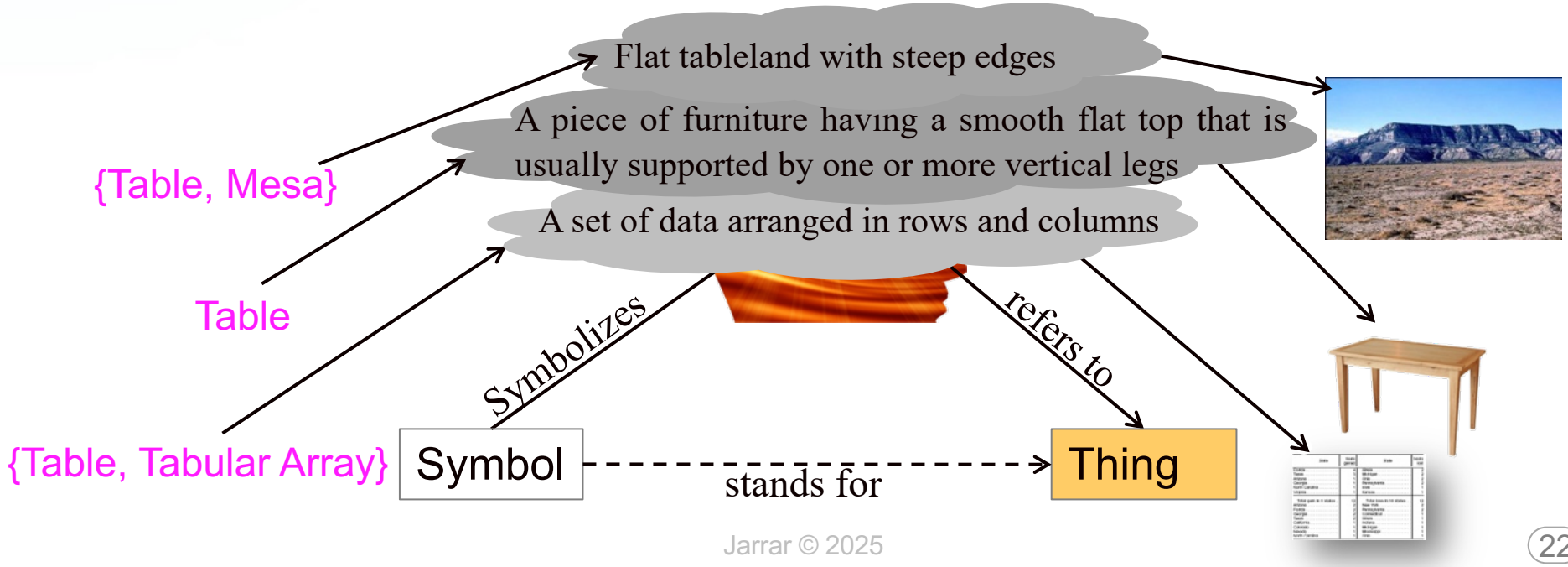
- **Polysemy**: is the capacity of a lexical unit to refer to multiple meanings/concepts. These meanings can be related or different.
- **Polysemy is the consequence of meaning evolution**. The constant discussion over how to name and what words mean is in the discourse of a community and implies language evolution. [T97]
- Note: the most frequent word forms are the most polysemous! [F]



Synonymy

الترادف

- **Synonymy**: different lexical units denoting the **same concept**
- Two lexical units are said to be **synonyms if they can be used interchangeably** in a certain context (/refer to the same extension).
- Mostly, synonyms are generated by the **parallel use**.
- Some lexicographers claim that no synonyms have exactly the same meaning (in all contexts or social levels of language)!!!!.



Synonymy

Synonymy in Wordnet

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).

{motorcar | machine | car | automobile | auto سيارة | مركبة}

He needs a car to get to work
He needs a motorcar to get to work
He needs an auto to get to work

Synonymy in Ontology Engineering

alternative labels/names of concepts

Definition: Synonymy Relation (see [J01])
Given two terms t_1 and t_2 lexicalizing concepts c_1 and c_2, respectively, then t_1 and t_2 are considered to be synonymous *iff* $c_1 = c_2$. In this way, synonymy can be defined as an equivalence relation $=_c$ between terms lexicalizing the same concept, thus it is a reflexive, symmetric and transitive relation.

Natural Language Processing

Lexical Semantics

In this lecture:

- Part 1: Linguistic Ontologies vs. Application Ontologies
- Part 2: What is Lexical Semantics
- Part 3: What is a Concept
- Part 4: Polysemy and Synonymy
- Part 5: **Multilingualism**
- Part 6: Distributional Semantics

Multilingually

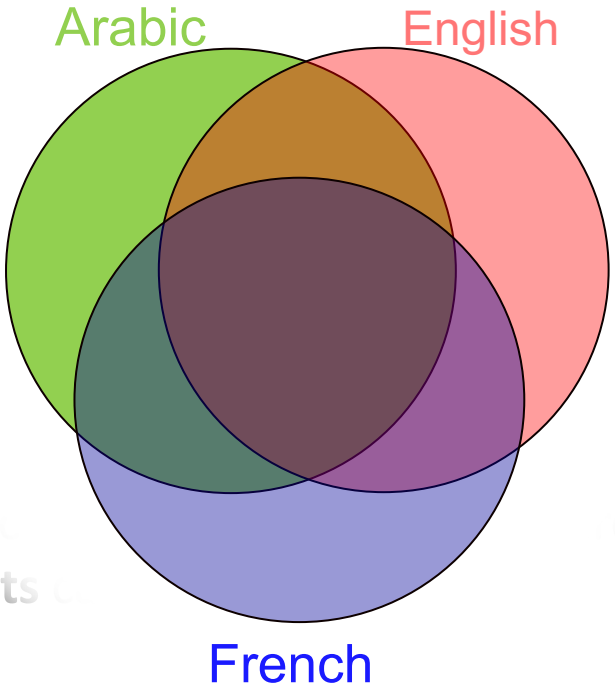
The language- independency of Concepts is problematic (See [J21]):

- Concepts/universals (that are benchmarked to scientific discoveries) **are language independent**.
- Concepts (that are benchmarked to *perceptions) are* **not totally language-independent**, as they typically depend on the perceptions and culture of the language-speakers.
- Many concepts are **shared cross languages**, especially if the speakers of these languages interact with each other.
- **The more interaction** between two communities speaking different languages, **the more shared concepts** can be found.

Multilingually

The language- independency of Concepts is problematic:

- Concepts/universals (that are benchmarked [illegible] discoveries) **are language independent**.
- Concepts (that are benchmarked [illegible] **language-independent**, as they typically de[illegible]re of the language-speakers.
- Many concepts are **shared cross**[illegible]s of these languages interact with ea[illegible]
- **The more interaction** between two [illegible]ent languages, **the more shared concepts** [illegible]



Natural Language Processing

Lexical Semantics

In this lecture:

- ❑ **Part 1:** Linguistic Ontologies vs. Application Ontologies
- ❑ **Part 2:** What is Lexical Semantics
- ❑ **Part 3:** What is a Concept
- ❑ **Part 4:** Polysemy and Synonymy
- ❑ **Part 5:** Multilingualism
- ❑ Part 6: **Distributional Semantics**

Distributional Semantics

الدلالة الاحصائية

Distributional Hypothesis:

Linguistic items with similar distributions have similar meanings

In other words, words that are used and occur in the same contexts tend to purport similar meanings (Harris 1954).

Example: “Car” and “Taxi”
“Solder and “Army”
“Boy” and “Gril”

Children can figure out how to use words they've rarely encountered before by generalizing about their use from distributions of similar words (Gleitman 2002).

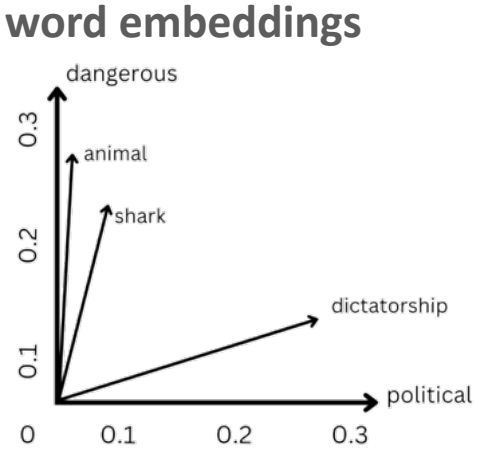
Distributional Semantics

الدلالة الاحصائية

Can we quantify semantic similarities between linguistic items based on their distributional properties in large samples of language data?

Distributional semantic similarity can be represented in different ways, including latent semantic analysis (LSA),Hyperspace Analogue to Language (HAL), syntax- or dependency-based models, random indexing, semantic folding and various variants of the topic model.

Semantic similarity between words using Word Embeddings

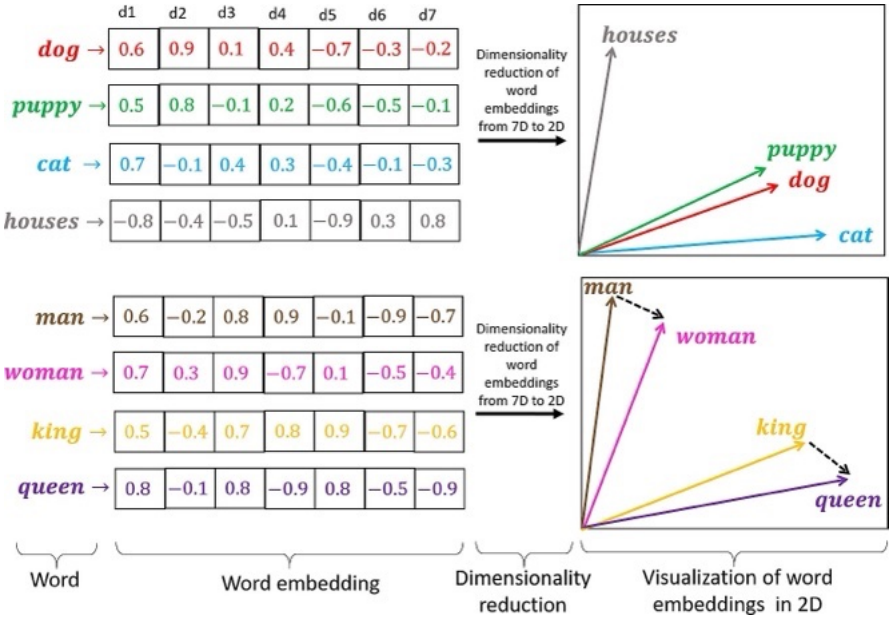


Word Embedding

الدلالة الاحصائية

Word Embedding:
A vector that encodes the meaning of the word in such a way that the words that are closer in the vector space are expected to be similar in meaning (Jurafsky et a., 2000).

can be obtained using language modeling and feature learning techniques (neural models), where words are mapped to vectors of real numbers.



Word Embedding

الدلالة الاحصائية

Word and phrase embeddings, when used as the underlying input representation, have been shown to boost the performance in NLP tasks, such as: syntactic parsing, **sentiment analysis**, **automatic creation of thesauri**, **word sense disambiguation**, **paraphrasing**, and others.

Underlying representation in:

- Word2Vec
- BERT
- GPT
- …

Read More: **A Compositional Distributional Model of Meaning**, by Stephen Clark Bob Coecke Mehrnoosh Sadrzadeh

https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=f86ef3e7b856d61ade62e643d87d288fef8827dd

References

[F] Christiane Fellbaum: **Lecture Notes on Words, Concepts, Meanings**

[T07] Rita Timmerman. **"Questioning the Univocity Ideal**. The Difference between Socio-cognitive Terminology and Traditional Terminology." *Journal of Linguistics* 18 (1997): 51-90.

[S04] Smith, B. (2004). **Beyond concepts: ontology as reality representation**. In Proceedings of the third international conference on formal ontology in information systems (pp.73-84).

[SCT04] Smith, B., Ceusters, W., Temmerman, R. (2004). **Wusteria**. Studies in health technology and informatics

[S06] Smith, B. (2006). **From concepts to clinical reality: an essay on the benchmarking of biomedical terminologies**. Journal of biomedical informatics, 39(3),

[W03] Wüster E. (2003): **The wording of the world presented graphically and terminologically.** Selected and translated by J.C. Sager (Lang.: eng). In: Terminology, 92, (pp.269-97).

[SO96] Sayyed Hossein Nasr and Oliver Leaman (1996), History of Islamic Philosophy, p. 315, Routledge,

[D92] Davidson, Herbert Alan (1992), *Alfarabi, Avicenna, and Averroes on Intellect: Their Cosmologies, Theories of the Active Intellect, and Theories of Human Intellect.*

[J21] Mustafa Jarrar: **The Arabic Ontology - An Arabic Wordnet with Ontologically Clean Content.** Applied Ontology Journal, 16:1, 1-26. IOS Press. 2021

[JH24] Mustafa Jarrar, Tymaa Hammouda: **Qabas: An Open-Source Arabic Lexicographic Database.** In Proceedings of LREC-COLING 2024, pages 13363–13370, Torino, Italia. ELRA and ICCL.

[JMHK2024] Mustafa Jarrar, Sanad Malaysha, Tymaa Hammouda, Mohammed Khalilia: **SALMA: Arabic Sense-Annotated Corpus and WSD Benchmarks.** Proceedings the 1st ArabicNLP, Part of the ACL 2023. ACL.

1. Sana Ghanem, Mustafa Jarrar, Radi Jarrar, Ibrahim Bounhas: **A Benchmark and Scoring Algorithm for Enriching Arabic Synonyms.** In Proceedings of GWC2023, (pp.274-283). Spain, 2023
2. Sanad Malaysha, Mustafa Jarrar, Mohammed Khalilia: **Context-Gloss Augmentation for Improving Arabic Target Sense Verification.** In Proceedings of GWC2023, (pp.274-283). Spain, 2023
3. Moustafa Al-Hajj, Mustafa Jarrar: **AraBglossBERT: Fine-Tuning BERT on Context-Gloss Pairs for WSD.** In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021). PP 40--48, 2021
4. Moustafa Al-Hajj, Mustafa Jarrar: **LU-BZU at SemEval-2021 Task 2: Word2Vec and Lemma2Vec performance in Arabic Word-in-Context disambiguation.** In Proceedings of the Fifteenth Workshop on Semantic Evaluation (SemEval2021) Task 2: Multilingual and Cross-lingual Word-in-Context Disambiguation (MCL-WiC). PP 748--755, Association for Computational Linguistics. 2021
5. Mustafa Jarrar, Eman Karajah, Muhammad Khalifa, Khaled Shaalan: **Extracting Synonyms from Bilingual Dictionaries.** The 11th International Global Wordnet Conference (GWC2021), Global Wordnet Association. (pp. 215-222). Pretoria, South Africa, 2021
6. Mustafa Jarrar, Hamzeh Amayreh: **An Arabic-Multilingual Database with a Lexicographic Search Engine.** The 24th International Conference on Applications of Natural Language to Information Systems (NLDB 2019). Pages(234-246). LNCS 11608, Springer. 2019
7. Mustafa Jarrar, Hamzeh Amayreh, John P. McCrae: **Representing Arabic Lexicons in Lemon - a Preliminary Study.** The 2nd Conference on Language, Data and Knowledge (LDK 2019). Pages(29-33). CEUR, Volume 2402. ISSN:1613-0073. Leipzig, Germany. 2019
8. Diana Alhafi, Anton Deik, Mustafa Jarrar: **Usability Evaluation of Lexicographic e-Services.** The 16th IEEE/ACS International Conference on Computer Systems and Applications (AICCSA). Pages(1-7). IEEE. Abu Dhabi, UAE. 2019
9. Mustafa Jarrar, Fadi Zaraket, Rami Asia, Hamzeh Amayreh: **Diacritic-Based Matching of Arabic Words.** ACM Asian and Low-Resource Language Information Processing. Volume 18, No 2, Pages(10:1-10:21), ACM, ISSN:2375-4699. December, 2018
10. Mustafa Jarrar, Werner Ceusters: **Classifying Processes and Basic Formal Ontology.** Proceedings of the 8th International Conference on Biomedical Ontology (ICBO 2017), Newcastle, UK. 2017
11. Mustafa Jarrar: **Building a Formal Arabic Ontology (Invited Paper)**. In proceedings of the Experts Meeting on Arabic Ontologies and Semantic Networks. Alecso, Arab League. Tunis, July 26-28, 2011.
12. Mustafa Jarrar: **Towards the notion of gloss, and the adoption of linguistic resources in formal ontology engineering.** In proceedings of the 15th International World Wide Web Conference (WWW2006). Edinburgh, Scotland. Pages 497-503. ACM Press. ISBN: 1595933239. May 2006.
13. Mustafa Jarrar, Anton Deik, Bilal Faraj: **Ontology-based Data and Process Governance Framework -The Case of e-Government Interoperability in Palestine**. Proceedings of the IFIP International Symposium on Data-Driven Process Discovery and Analysis (SIMPDA'11). Pages(83-98). 2011.
14. Mustafa Jarrar and Robert Meersman: **Ontology Engineering -The DOGMA Approach**. Book Chapter in "Advances in Web Semantics I". Chapter 3. Pages 7-34. LNCS 4891, Springer.ISBN:978-3540897835. (2008).