Natural Language Processing

Wordnets

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:
Arabic, Arabic NLP, NLP, Natural Language Processing, Thesauri, linguistic ontology, Arabic Ontology, Wordnet, Arabic WordNet, EURO WordNet, Global WordNet, Lexical Semantics, Word sense disambiguation, mental lexicon, Synset, Concept, Gloss, Polysemy, Semantic Relations, Hyponymy, Meronymy, Antonymy, Ontology, شبكة المفردات، وردنت، حوسبة الدلالة، معجم ذهني، مترادفات، مفهوم، تعريف، مكنز، حوسبة اللغة، الانطولوجيا العربية، أنطولوجيا لغوية، تعدد اللغات، الترادف اللغوي، تعدد المعاني، التضاد، تصنيف المعاني، علاقات جزء-كل

Natural Language Processing

Introduction to Wordnets

In this lecture:

- Part 1: **What and why Thesauri**
- Part 2: WordNet
- Part 3: EuroWordnet
- Part 4: Global Wordnet
- Part 5: Discussion
- Part 6: Practice

Reading

Everything in these slides + everything I say

[MBC93] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller: **Introduction to WordNet: An On-line Lexical Database**. International Journal of Lexicography, Vol. 3, Nr. 4. Pages 235-244. (1990) http://wordnetcode.princeton.edu/5papers.pdf

[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

Why Lexical Semantic Resources?

The importance of lexical semantic resources (such as thesauri, wordnets, linguistic ontologies) is increasing in many application areas, such as:

- Word sense disambiguation
- Smarter Information search and retrieval
- Multilingual semantic web and knowledge graphs
- NLP tasks and applications (classification/ summarization/translation)
- Data integration
- Multilingual big data
- among many others.

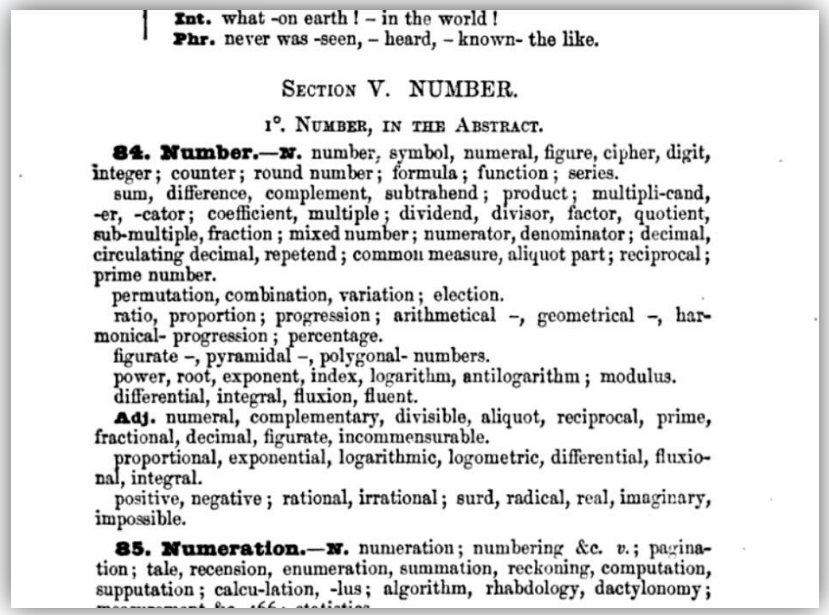
Thesaurus (مكنز) as a source of semantics

A **list of words classified as near-synonyms**.

Examples:

Thesaurus of English words and phrases

Peter Mark Roget · 1883



Int. what -on earth ! – in the world !
Phr. never was -seen, – heard, – known- the like.

SECTION V. NUMBER.

1°. NUMBER, IN THE ABSTRACT.

84. Number.—N. number, symbol, numeral, figure, cipher, digit, integer; counter; round number; formula; function; series.
sum, difference, complement, subtrahend; product; multipli-cand, -er, -cator; coefficient, multiple; dividend, divisor, factor, quotient, sub-multiple, fraction; mixed number; numerator, denominator; decimal, circulating decimal, repetend; common measure, aliquot part; reciprocal; prime number.
permutation, combination, variation; election.
ratio, proportion; progression; arithmetical –, geometrical –, harmonical- progression; percentage.
figurate –, pyramidal –, polygonal- numbers.
power, root, exponent, index, logarithm, antilogarithm; modulus.
differential, integral, fluxion, fluent.
Adj. numeral, complementary, divisible, aliquot, reciprocal, prime, fractional, decimal, figurate, incommensurable.
proportional, exponential, logarithmic, logometric, differential, fluxional, integral.
positive, negative; rational, irrational; surd, radical, real, imaginary, impossible.

85. Numeration.—N. numeration; numbering &c. *v.*; pagination; tale, recension, enumeration, summation, reckoning, computation, supputation; calcu-lation, -lus; algorithm, rhabdology, dactylonomy;

المكنز العربي المعاصر

محمود إسماعيل صيني، ناصيف مصطفى عبدالعزيز، مصطفى أحمد سليمان 1993

ا ج ت ٤

اجْتَهَدَ: ج ه د. ف. ثابَرَ، جَدَّ، واظَبَ.
اجْتِياح: ج و ح. س. إهْلاك، اسْتِئْصال، دَمار، تَدْمير.
أَجْدَبَ: ج د ب. ف. أمْحَلَ، أقْحَط، أسْنَتَ.
أَجْدَرُ: ج د ر. ص. أحَقُّ، أحْرى، أوْلى، قَمِنٌ، خَليق، أفْضَل.
أَجْر: أ ج ر. س. ثَواب، جَزاء.
أُجْرَة: أ ج ر. س. كِراء، جَزاء، مُكافَأة، ثَواب، مَثُوبة، خَرْج، جُعْل.
أَجَلّ: أ ج ل. ف. أخَّر، نَفَّس، أمْهَل، أنْظَر، أرْجَأ، أنْسَأ، أكْلَأ، نَجَّم.
أَجَلّ: ج ل ل. ف. عَظَّم، وقَّر، احْتَرَم، بَجَّل.
أَجَمَة: أ ج م. س. غابَة، مَقْصَبة، أيْكَة، غَيْضَة، دَغْل، عَرِينَة، حَرَجَة، رَحَبَة، حديقة.
أَجْوَد: ج و د. ص. أفْضَل، خَيْر، أنْفَع، أجْدى، أرْبَح، أوْفَر، أوْفى، أجْزَل، أزْكى، أعود.
أحاط ١: ح ي ط. ف. (بـ) أحاق، اكْتَنَف، الْتَفَّ، طاف.
أحاط ٢: ح ي ط. ف. (بـ) أحْدَق بـ، سَيْطَر على، اسْتَوْلى على، تَمَكَّن من.
احاطة: ح ي ط. س. إحْداق، حَقّ، إحاقة، اكْتِناف، اسْتِدارَة، الْتِفاف، إطافَة.
أَحَبَّ: ح ب ب. ف. وَدَّ، هَوى، مَقَ، شَغَفَ بـ، أُغْرِم بـ، أُولِعَ بـ، عَشِقَ، كَلِفَ بـ، هام، وَلِه.
أحْبَطَ: ح ب ط. ف. أبْطَل، أتْلَف، أفْسَد، ألْغى.
احْتَجَّ ١: ح ج ج. ف. عارَض، اعْتَرَض، اسْتَنْكَر، صَدَّ.
احْتَجَّ ٢: ح ج ج. ف. اسْتَنْهَد، تَذَرَّع.
احْتِجاج: ح ج ج. س. اعْتِراض، مُعارَضَة، اسْتِياء، تَبَرُّم، شَجْب.
احْتَجَبَ: ح ج ب. ف. اسْتَتَر، اخْتَفى، تَوارَى، أفَلَ، غاب.
احْتَجَزَ ١: ح ج ز. ف. اقْتَطَع، سَرَق.

Thesaurus (مكنز) as a source of semantics

A **list of words classified as near-synonyms**;

or

it can be seen as pairs of terms connected through "*RelatedTo*" and/or a "*Broader/Narrow*" relations.

However, such relations are **semantically-poor** and imprecise relationships between words and not sufficient for most IT-based applications.

➔ **From thesaurus to wordnet**

Natural Language Processing

Introduction to Wordnets

In this lecture:

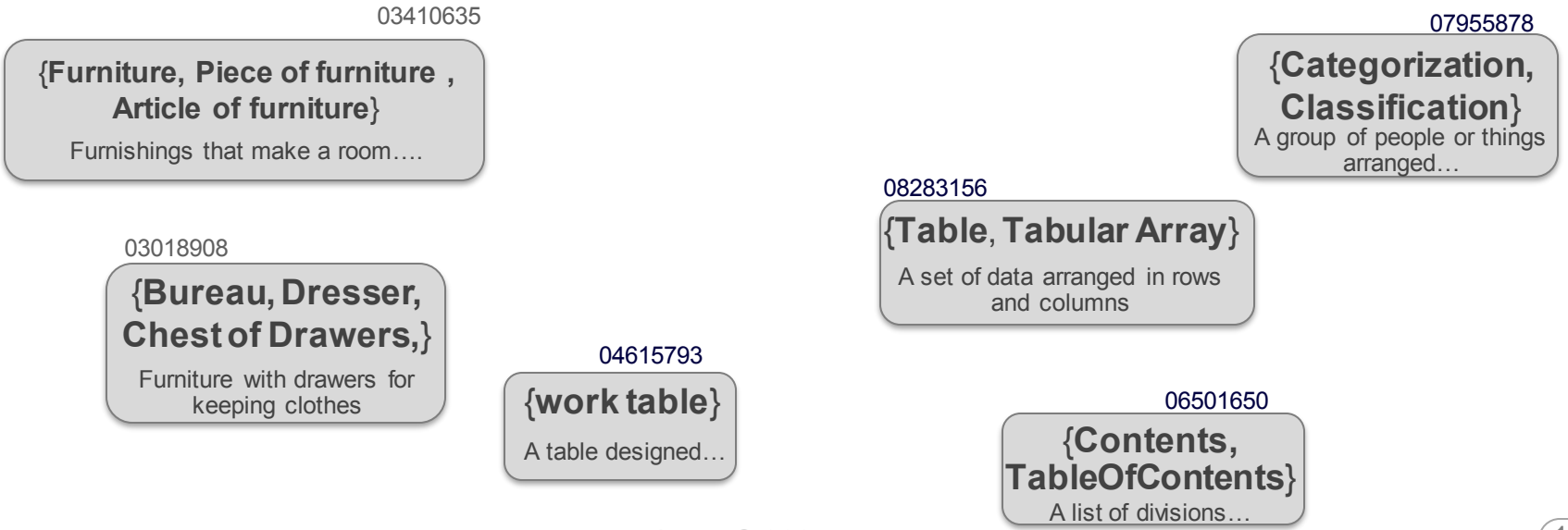
- ❑ Part 1: What and why Thesauri
- ❑ Part 2: **WordNet**
- ❑ Part 3: EuroWordnet
- ❑ Part 4: Global Wordnet
- ❑ Part 5: Discussion
- ❑ Part 6: Practice

What is WordNet?
شبكة مفردات

- In 1985 a group of **psychologists and linguists** at Princeton University started to develop a "**mental lexicon**" معجم ذهني .
- You may also call it: electronic dictionary, mental dictionary, **semantic Network**, hyperdimensional thesaurus, lexicographic database, (recently called linguistic ontology). etc.
- Includes **most frequent words** (nouns, adjectives, adverbs, verbs).
- **Organized by meaning**: words in close proximity are semantically similar.
- Can be used by humans and machines.
- Human users and computers can browse WordNet and find words that are meaningfully related to their queries.
- **Available online**, for downloading! http://wordnet.princeton.edu

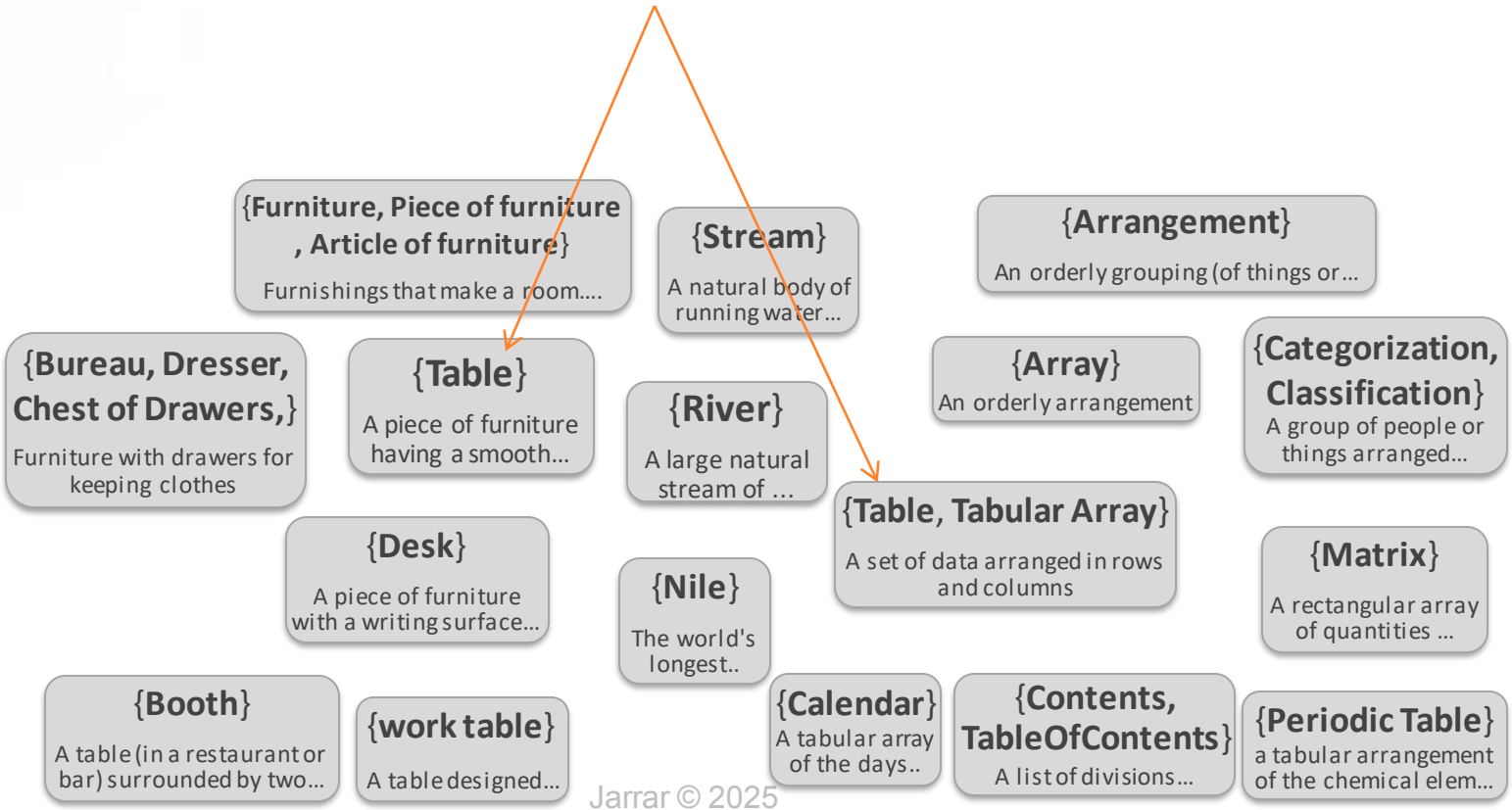
Basic Concepts

- English words are grouped into sets of synonyms called a **Synset**.
- Each synset is given a unique **SynsetID**.
- Each synset *signify* that a **Concept** exist – expressing a meaning.
- Each word form-meaning pair is unique: **Sense**.
- Each synset is described by a **gloss** (examples of contexts).



WordNet: **Polysemy**

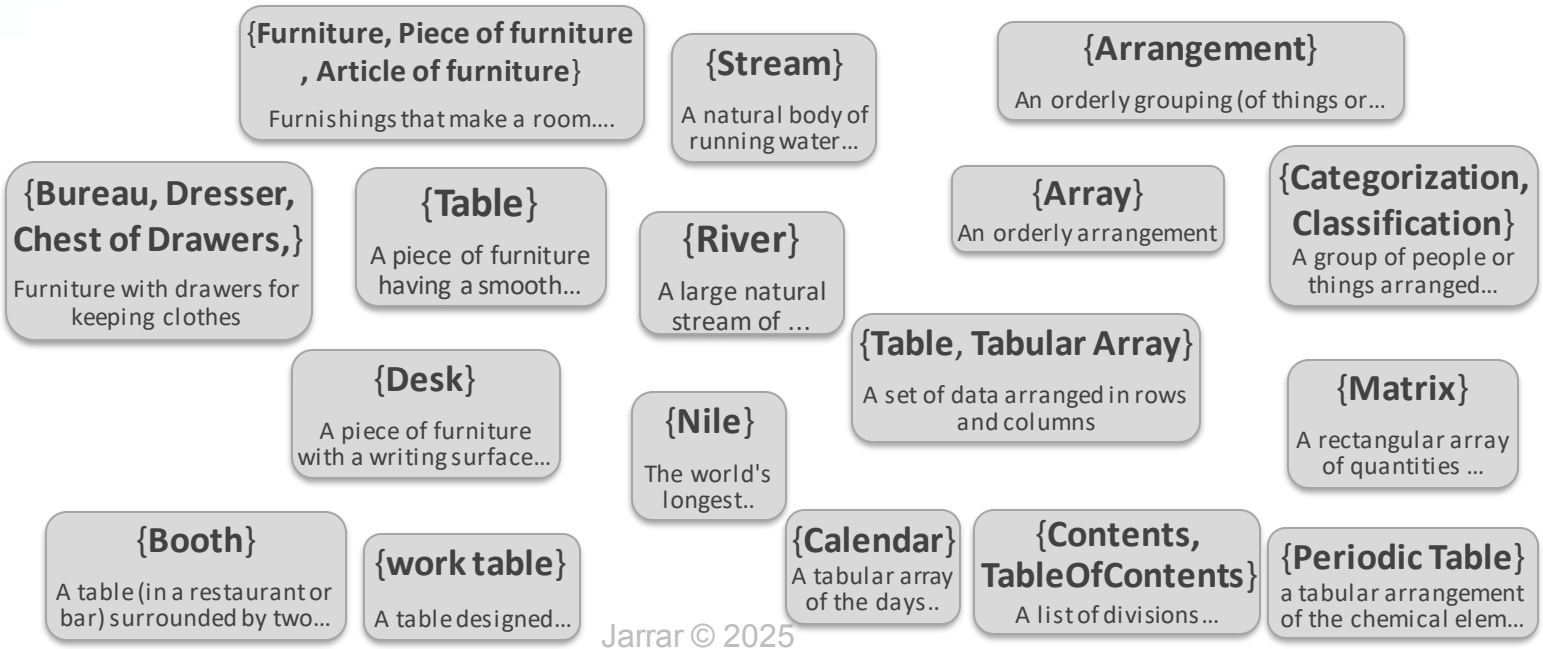
- A word that appears in *n* synsets is *n*-fold polysemous.
- For example: “Table” here is two-fold polysemous



WordNet: **Glosses**

A short gloss is provided for each sysnet.

Glosses are examples of contexts for many word-sense pairs, telling us how words with specific senses are being used in context.



WordNet: **Glosses**

155 287 word forms, groups into

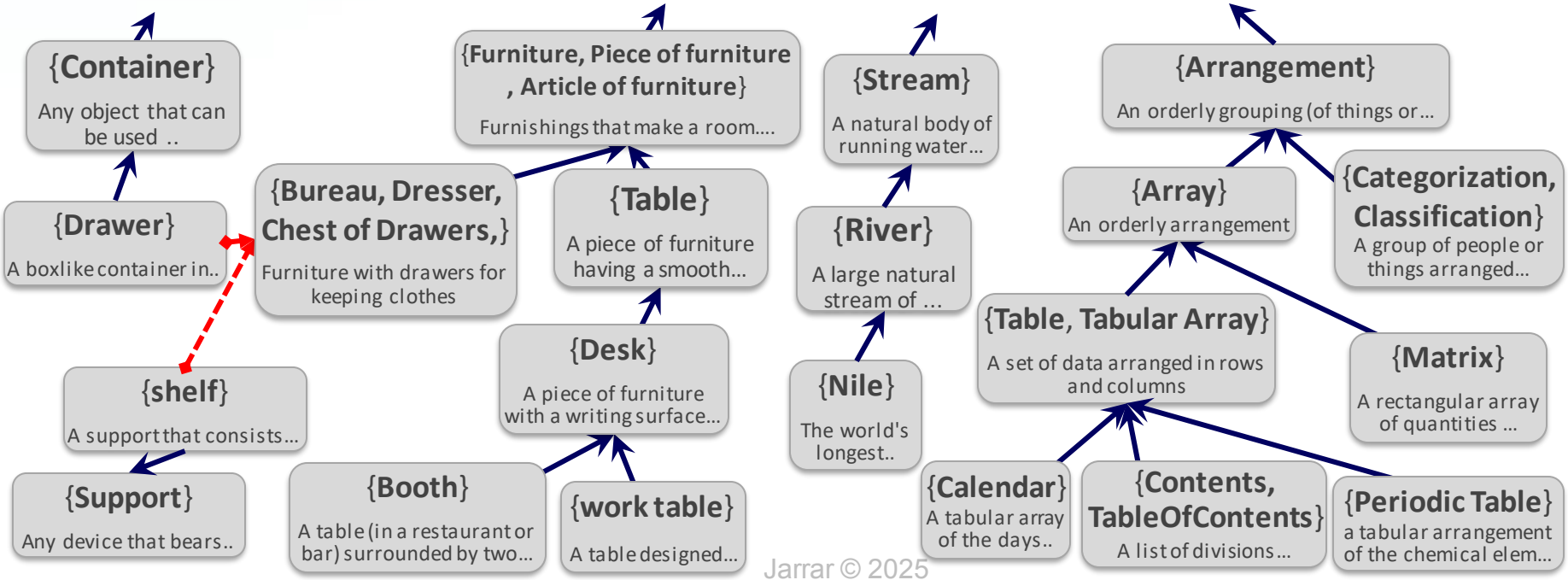
117 659 synsets

	WordForms	Synsets
noun	117,798	82,115
verb	11,529	13,767
adjective	21,479	18,156
adverb	4,481	3,621
Total	**155,287**	**117,659**

WordNet Semantic Relations

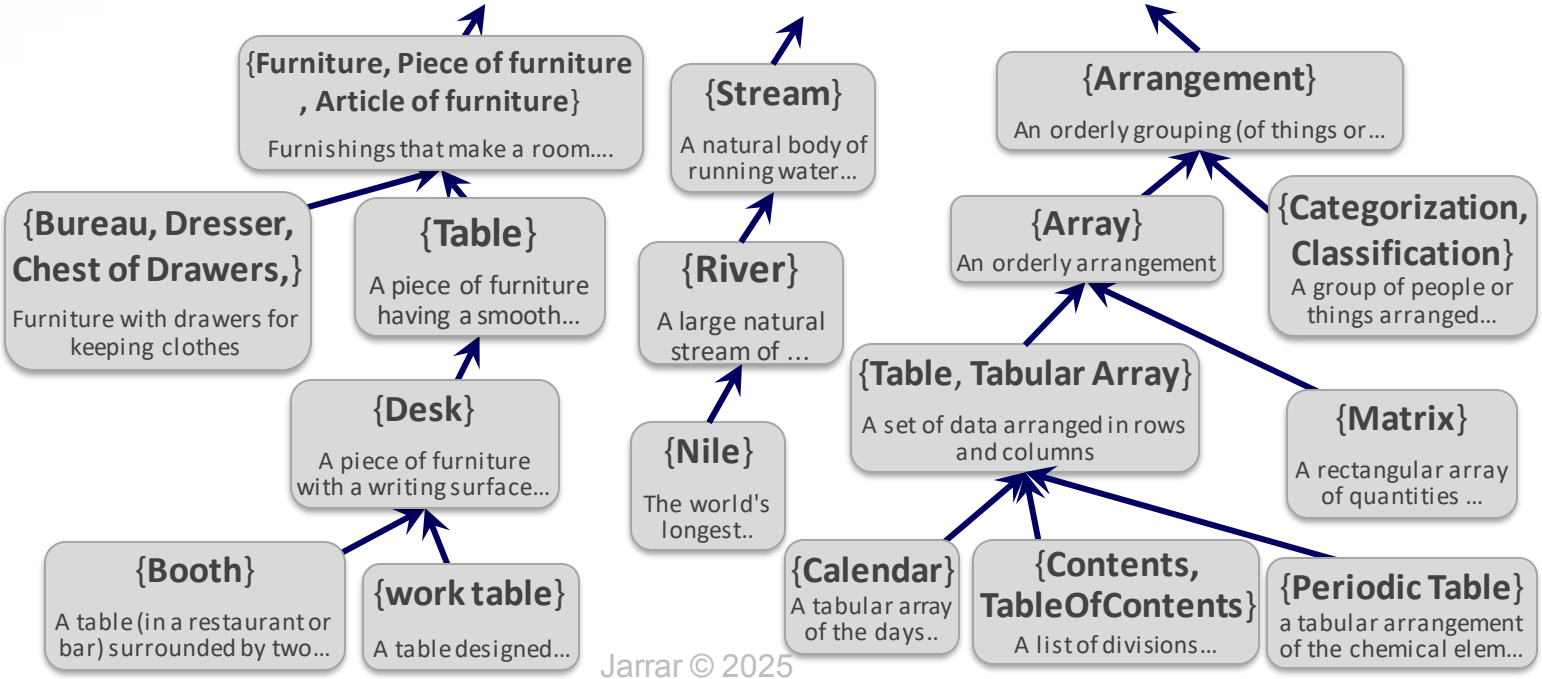
Synsets are interconnected with semantic relations, forming a large semantic network (graph). Such Relations are:

- **Hyponymy**, also called “Is a” relation, or sub/superordinate.
- **Meronymy**, also called “part of” relation



WordNet Relations: **Hyponymy**

- A synset {x, x′, . . .} is hyponym of the synset {y, y′, . . .} **if native English speakers accept sentences like x is a (kind of) y**. E. g., *Table/Tabular Array* is a kind of *Array*, *Array* is a kind of *Arrangement*,…
- Hyponymy is transitive and asymmetrical. So as Hyponymy generates a hierarchical semantic structure, a hyponym inherits all the features of the more generic concept and adds at least one feature that distinguishes it from its superordinate.



WordNet Relations: **Hyponymy**

- A synset {x, x′, . . .} is hyponym of the synset {y, y′, . . .} **if native English speakers accept sentences like x is a (kind of) y**. E. g., *Table/Tabular Array* is a kind of *Array*, *Array* is a kind of *Arrangement*,…
- Hyponymy is transitive and asymmetrical. So as Hyponymy generates a hierarchical semantic structure, a hyponym inherits all the features of the more generic concept and adds at least one feature that distinguishes it from its superordinate.

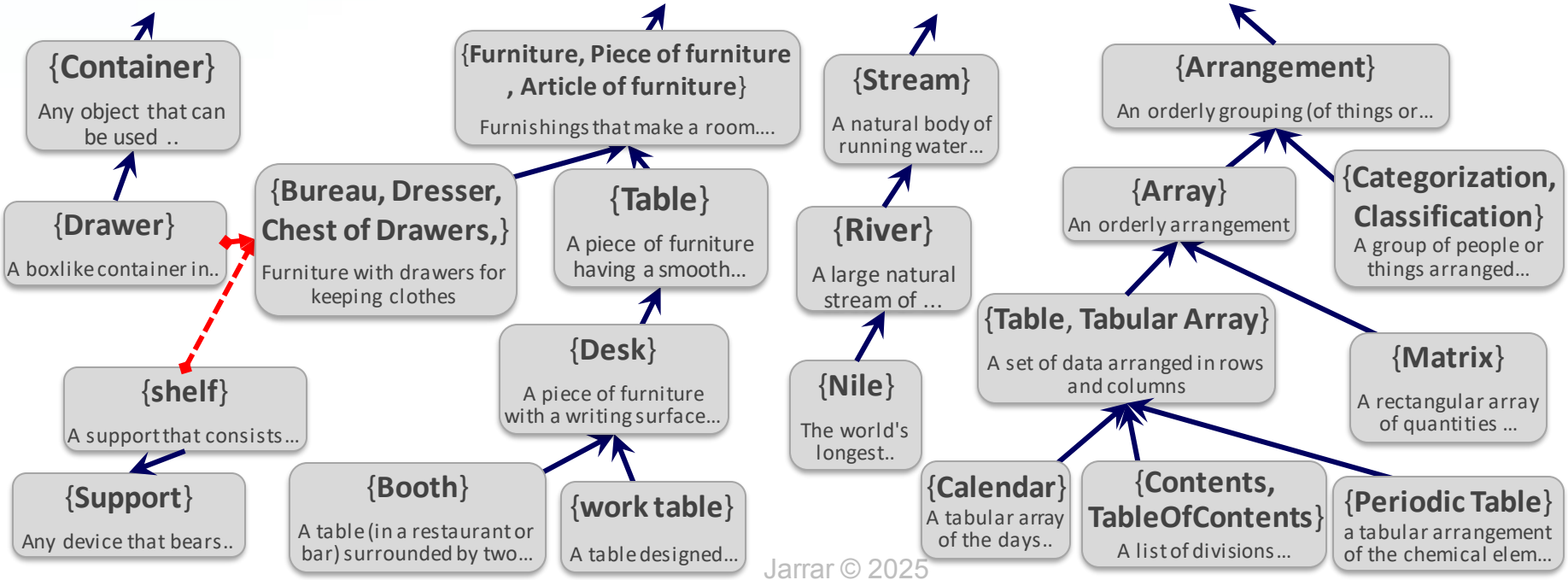
The WordNet hierarchy is about 16 levels

Top Level Nouns (25 unique beginners)

{act, action, activity}	{natural object }
{animal, fauna}	{natural phenomenon }
{artifact }	{person, human being}
{attribute, property }	{plant, flora}
{body, corpus}	{possession}
{cognition, knowledge}	{process}
{communication}	{quantity, amount}
{event, happening}	{relation }
{feeling, emotion}	{shape}
{food}	{state, condition}
{group, collection}	{substance}
{location, place }	{time}
{motive}	

WordNet Relations: **Meronymy**

- A synset $\{x, x', \ldots\}$ is meronym of the synset $\{y, y', \ldots\}$ if native English speakers accept sentences like y has an x (as a part) or An x is a part of y. E. g., *Finger* is part of *Hand* , *Hand* is part of *Arm*, *Arm* is part of *Body*.
- Meronymy is transitive (with qualification) and asymmetrical relations, and forms a part hierarchy.
- Synsets may have multiple hypernyms



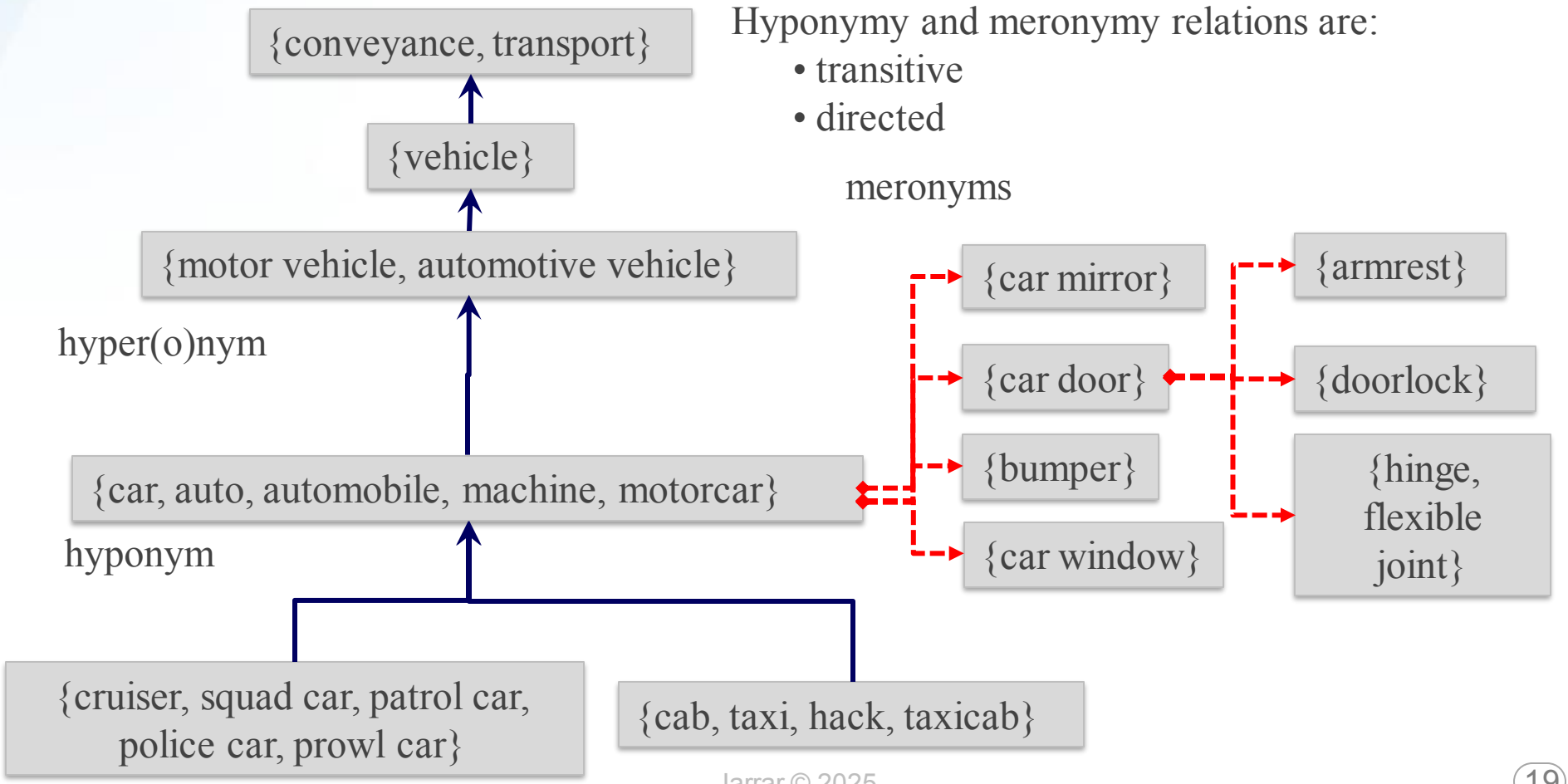
Exercise

Find the hyponyms and meronyms of this synset

{car, auto, automobile, machine, motorcar}

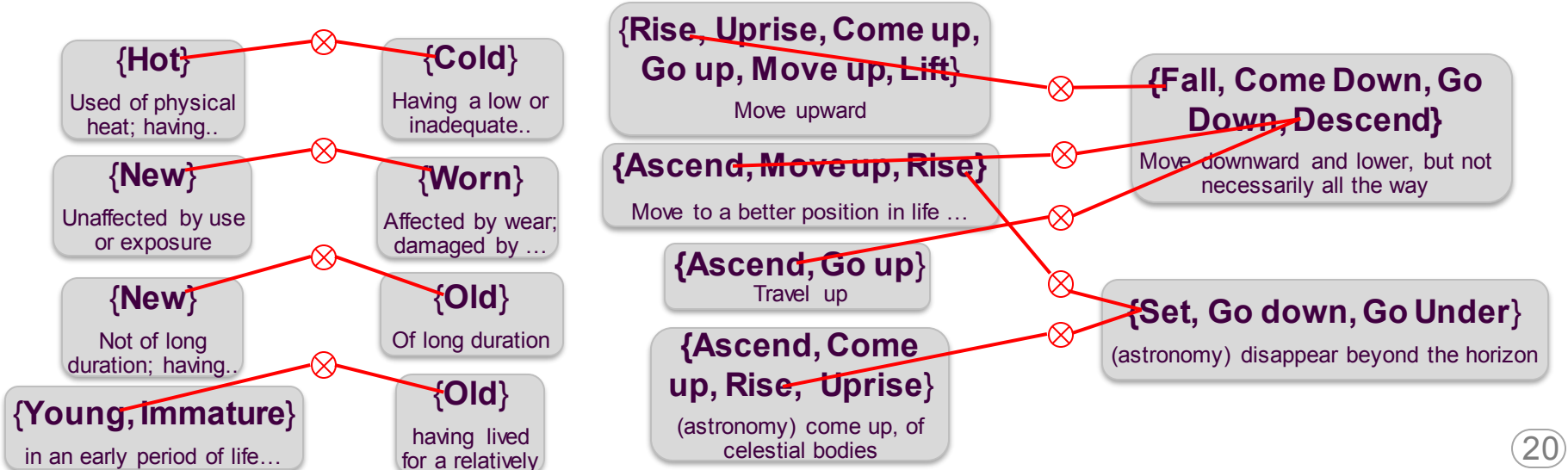
WordNet Relations: **Another Example**

[Vossen]



WordNet Relations: **Antonymy**

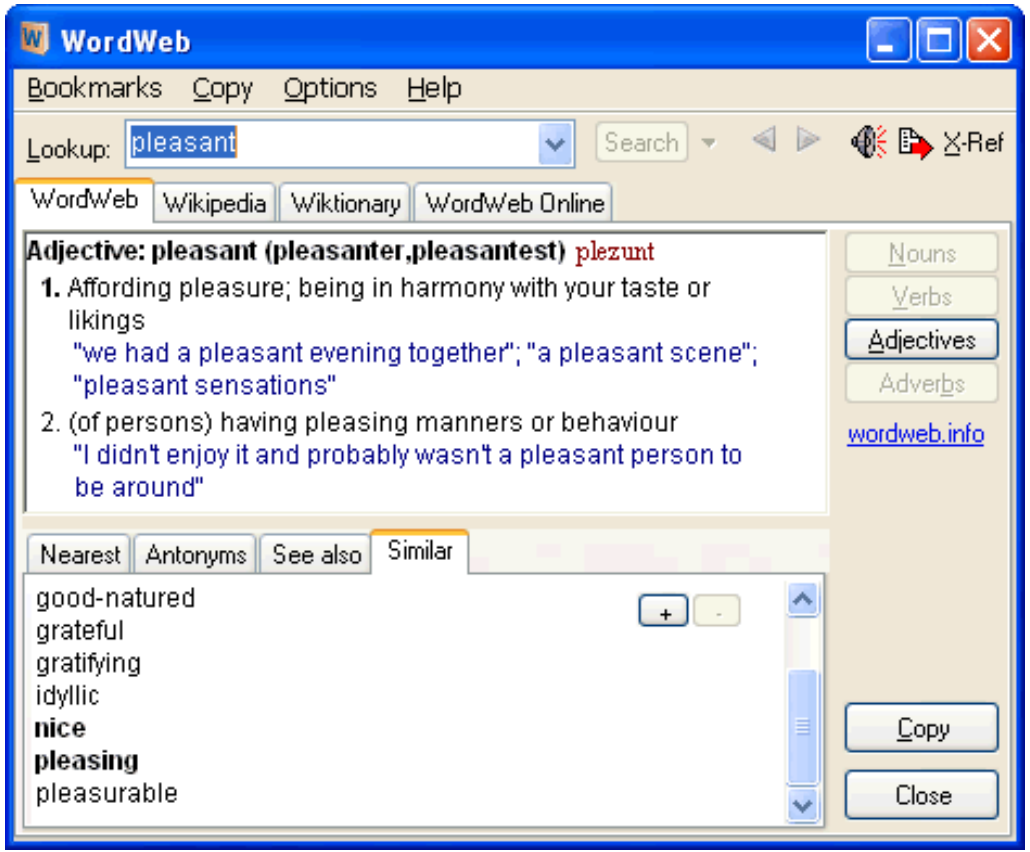
- The antonym of a word x is sometimes not-x, but not always. For example, *rich* and *poor* are antonyms, but to say that someone is not rich does not imply that they must be poor; many people consider themselves neither rich nor poor.
- Antonymy, which seems to be a simple symmetric relation, is actually quite complex, yet speakers of English have little difficulty recognizing antonyms when they see them. For example, the meanings {rise, ascend } and {fall, descend} may be conceptual opposites, but they are not antonyms; [rise/fall] are antonyms and so are [ascend/descend], but most people hesitate and look thoughtful when asked if rise and descend, or ascend and fall, are antonyms
- Antonymy is a lexical relation between word forms, not a semantic relation between word meanings. Or, some call it semantic relations between words [MPC93].



WordWeb

http://wordweb.info/free/

A nice and intuitive interface for WordNet



Other WordNet Relations

- Although the main interest of WordNet was on specifying semantic relations but other lexical/morphological relations between word forms were added.

- For example: stems, singular-plural, verb tenses, etc.

Is WordNet a Thesaurus?

Yes:

- it groups together meaningfully related words

and more:

- WordNet provides **more accurate** relations,

  Thesaurus contains only related-to.

- Related words are **linked to specific concepts** (disambiguated),

  Thesaurus is a “bag of words”

➔ **Wordnets are next generation Thesauri**

Is WordNet an Ontology?

Ontological Precision:

WordNet: based on what native speakers agree roughly.

Ontology: based on Scientific and philosophical findings.

Classification:

WordNet: based on what native speakers agree roughly (Student IsA person)

Ontology: based on strict formal methodologies (student IsA role)

Formal Specification:

WordNet: logically vague (and, contains concepts without instance)

Ontology: strictly formal (every concepts can be instantiated)

Examples of ontological matters in WordNet

Examples problems in WordNet, which limited its use in IT applications:

- (Nile *Is-a* River) is **formal mistake**, Nile is an instance of River.
- (Student *Is-a* Person) is **ontologically incorrect**; Student is a *Role*
- (Italy *Is-a* Land5) and (Italy *Is-a* Nation) is **ontologically incorrect**. cannot subsume the two disjoint concepts, land5 and nation, at the same time.

- (Reflate$_2$ *Is-a* Inflate$_3$) (Inflate$_3$ *Is-a* Change$_1$) and (Reflate$_2$ *Is-a* Change$_1$) is **meaningless**, this is an implied relation.
- (Restrain$_1$ *Is-a* Inhibit$_4$) and (Inhibit$_4$ *Is-a* Restrain$_1$) is a **cycle**.
- (Islamic Month *Is-a* Month) is **inaccurate**, Month = twelve divisions of the Gregorian year (i.e., 30.43 days); but Islamic month is 29.53 days.
- Moring and Evening Stars as different stars is **inaccurate**. They are the same instance (i.e., Venus) that people see at different occasions.

➔ **From thesaurus to wordnet to linguistic ontology**

Natural Language Processing

Introduction to Wordnets

In this lecture:

- ❑ Part 1: What and why Thesauri
- ❑ Part 2: What is WordNet
- ❑ Part 3: **EuroWordnet**
- ❑ Part 4: Global Wordnet
- ❑ Part 5: Discussion
- ❑ Part 6: Practice

EURO WordNet

[Vossen]

- The development of a multilingual database with WordNets for several European languages.

- Funded by the European Commission, DG XIII, LE2-4003 and LE4-8328
- March 1996 - September 1999 (2.5 Million EURO)
  http://www.hum.uva.nl/~ewn
  http://www.illc.uva.nl/EuroWordNet/finalresults-ewn.html

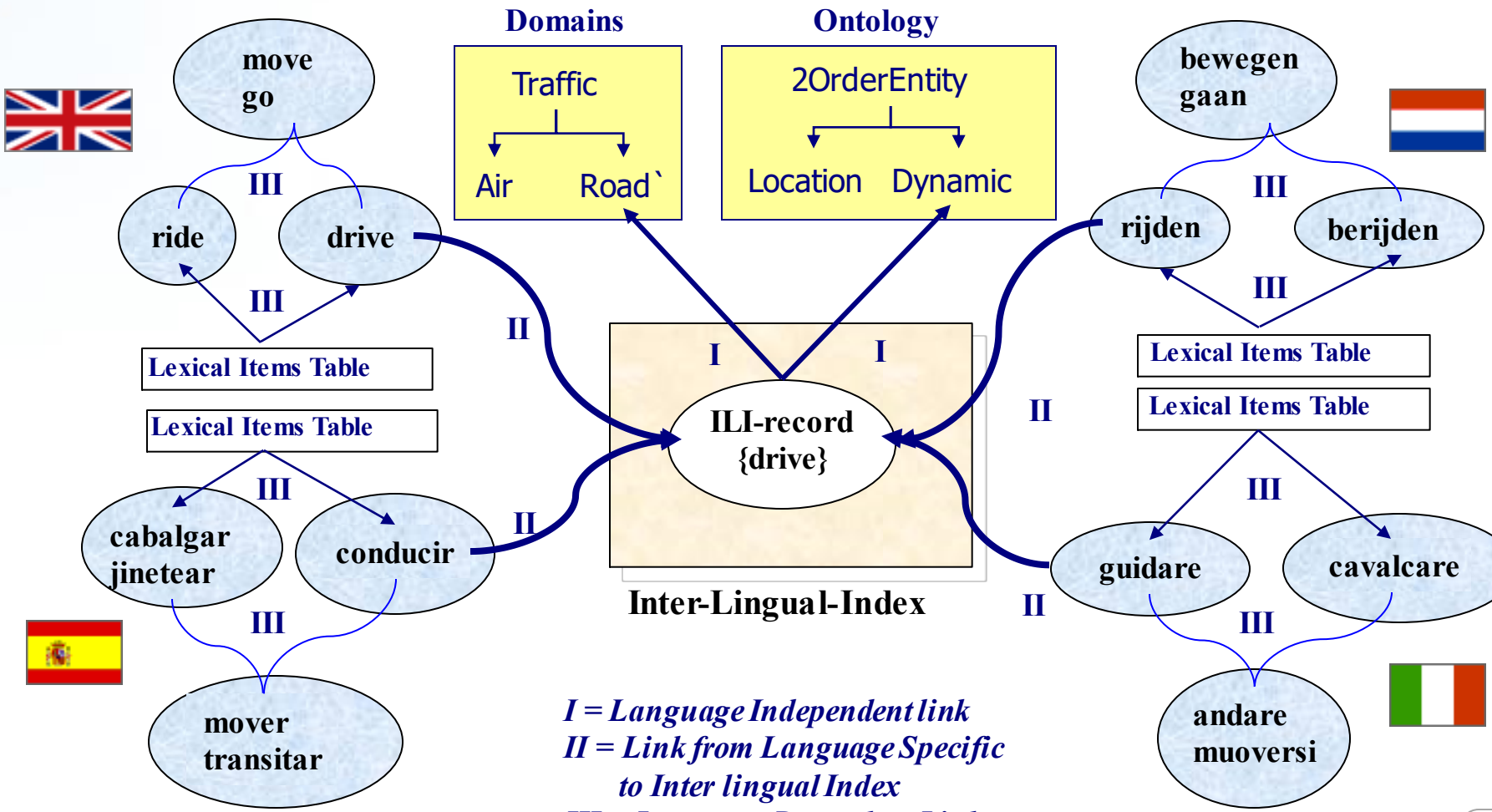
- **Languages covered:**
  EuroWordNet-1 (LE2-4003): English, Dutch, Spanish, Italian
  EuroWordNet-2 (LE4-8328): German, French, Czech, Estonian.

- **Size of vocabulary:**
  EuroWordNet-1: 30,000 concepts - 50,000 word meanings.
  EuroWordNet-2: 15,000 concepts- 25,000 word meaning.

- **Type of vocabulary:**
  the most frequent words of the languages
  all concepts needed to relate more specific concepts.

EURO WordNet Model

[Vossen]



The Multilingual Design

[Vossen]

- Inter-Lingual-Index: **unstructured fund of concepts** to provide an efficient mapping across the languages;
- Index-records are mainly **based on WordNet synsets** and consist of synonyms, glosses and source references;
- Various types of **complex equivalence relations** are distinguished;
- Equivalence relations from synsets to index records: **not on a word-to-word basis;**
- **Indirect** matching of synsets linked to the same index items;

EURO WordNet Model

[Vossen]

- WordNets are unique language-specific structures:
  - same organizational principles: synset structure and same set of semantic relations.
  - different lexicalizations
  - differences in synonymy and homonymy:

    "decoration" in English versus "versiersel/versiering" in Dutch

    "bank" in English (money/river) versus "bank" in Dutch (money/furniture)

•BUT also different relations for similar synsets

Some Downsides of the EuroWordNet Model

[Vossen]

- Construction is not done uniformly
- Coverage differs
- Not all wordnets can communicate with one another, i.e. linked to different versions of English wordnet
- Proprietary rights restrict free access and usage
- A lot of semantics is duplicated
- Complex and obscure equivalence relations due to linguistic differences between English and other languages

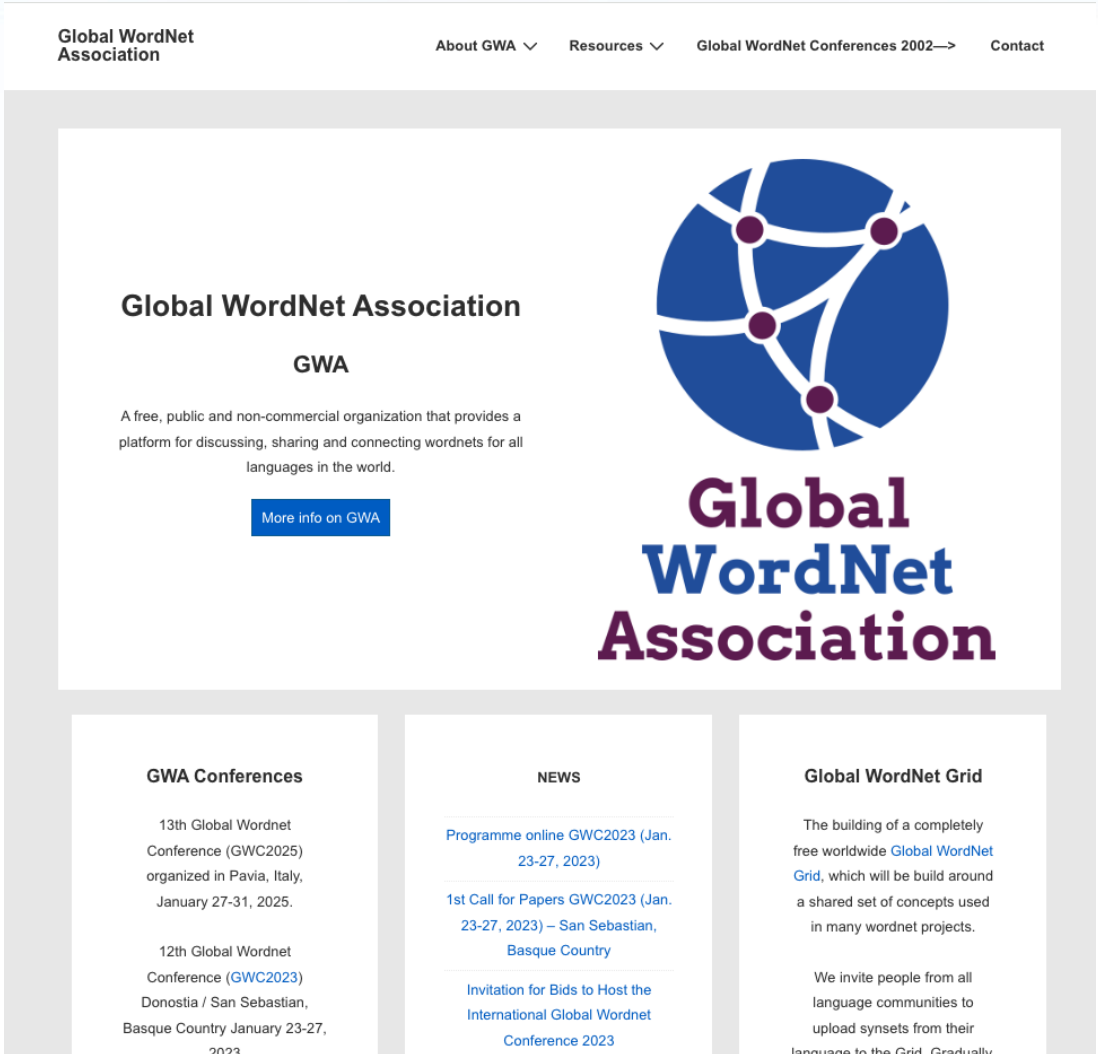
Natural Language Processing

Introduction to Wordnets

In this lecture:

- ❑ Part 1: What and why Thesauri
- ❑ Part 2: What is WordNet
- ❑ Part 3: EuroWordnet
- ❑ Part 4: **Global Wordnet**
- ❑ Part 5: Discussion
- ❑ Part 6: Practice

From EuroWordNet to Global WordNet



http://www.globalwordnet.org

From EuroWordNet to Global WordNet

[Vossen]

- EuroWordNet ended in 1999
- Global Wordnet Association was founded in 2000 to maintain the framework: http://www.globalwordnet.org
- Currently, wordnets exist for more than 80 languages, including:

  Albanian, Arabic, Bantu, Basque, Chinese, Bulgarian, Estonian, Hebrew, Icelandic, Japanese, Kannada, Korean, Latvian, Nepali, Persian, Romanian, Sanskrit, Tamil, Thai, Turkish, Zulu...

- Many languages are genetically and typologically unrelated

➔ The Arabic WordNet extension was not successful.

Arabic WordNet

- Literal and ad hoc translation for 10000 English synsets, and never extended!
- The 10000 synsets were selected as the following:
  - A set of concepts (called **base concepts**) were selected as they exist in 12 languages (in EuroWordNet and BalkeNet, (Elkateb et al 2006), thus they are assumed to also exist in Arabic.
  - The base concepts were then extended mostly downwards with more specific concepts, and upwards with more general concepts, to improve the maximal connectivity of those base concepts.
  - Some new synsets were added – mostly names of cities

Download: https://sourceforge.net/projects/awnbrowser

Natural Language Processing

Introduction to Wordnets

In this lecture:

- Part 1: What and why Thesauri
- Part 2: What is WordNet
- Part 3: EuroWordnet
- Part 4: Global Wordnet
- Part 5: **Discussion**
- Part 6: Practice

Natural Language Processing

Introduction to Wordnets

In this lecture:

- ❑ Part 1: What and why Thesauri
- ❑ Part 2: What is WordNet
- ❑ Part 3: EuroWordnet
- ❑ Part 4: Global Wordnet
- ❑ Part 5: Discussion
- ❑ Part 6: **Practice**

References

[Vossen] Piek Vossen, **From WordNet,to EuroWordNet,to the Global Wordnet Grid**. Lecture notes

[MBC93] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller: **Introduction to WordNet: An On-line Lexical Database**. International Journal of Lexicography, Vol. 3, Nr. 4. Pages 235-244. (1990)

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

1. 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.
2. 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.
3. 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
4. Sanad Malaysha, Mustafa Jarrar, Mohammed Khalilia: **Context-Gloss Augmentation for Improving Arabic Target Sense Verification.** In Proceedings of GWC2023, (pp.274-283). Spain, 2023
5. 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
6. 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
7. 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
8. 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
9. 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
10. 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
11. 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
12. Mustafa Jarrar, Werner Ceusters: **Classifying Processes and Basic Formal Ontology.** Proceedings of the 8th International Conference on Biomedical Ontology (ICBO 2017), Newcastle, UK. 2017
13. 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.
14. 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.
15. 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.
16. 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. (2008).
17. Mustafa Jarrar: **Tutorial on Arabic Ontology Engineering**. The ACS/IEEE International Conference on Computer Systems and Applications. Tunis, 2017
18. Mustafa Jarrar, Maria Keet, and Paolo Dongilli: **Multilingual verbalization of ORM conceptual models and axiomatized ontologies** . Technical report. STARLab, Vrije Universiteit Brussel, February 2006.
19. Mustafa Jarrar: **Mapping ORM into the SHOIN/OWL Description Logic- Towards a Methodological and Expressive Graphical Notation for Ontology Engineering** . In OTM 2007 workshops: Proceedings of the International Workshop on Object-Role Modeling (ORM'07). Pages (729-741), LNCS 4805, Springer. ISBN:9783540768890. Portogal. November, 2007