

Introduction to **Wordnets**

Mustafa Jarrar

[Birzeit University](#)



Watch this lecture and download the slides



Course Page: <http://www.jarrar.info/courses/WordnetBasics.pdf>

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

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>

[GGO02] Aldo Gangemi , Nicola Guarino , Alessandro Oltramari , Ro Oltramari , Stefano Borgo: **Cleaning-up WordNet's Top-Level**. In Proc. of the 1st International WordNetConference (2002)
<http://citeseer.ist.psu.edu/viewdoc/download;jsessionid=C9962DFEDD793F3F839426B774BC9BAF?doi=10.1.1.11.4064&rep=rep1&type=pdf>

Introduction to Wordnets

In this lecture:



- ☐ Part 1: **What and why Thesauri**
- ☐ Part 2: What is WordNet
- ☐ Part 3: EuroWordnet
- ☐ Part 4: Global Wordnet

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

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**

Introduction to Wordnets

In this lecture:

☐ Part 1: What and why Thesaurus



☐ Part 2: **What is WordNet**

☐ Part 3: EuroWordnet

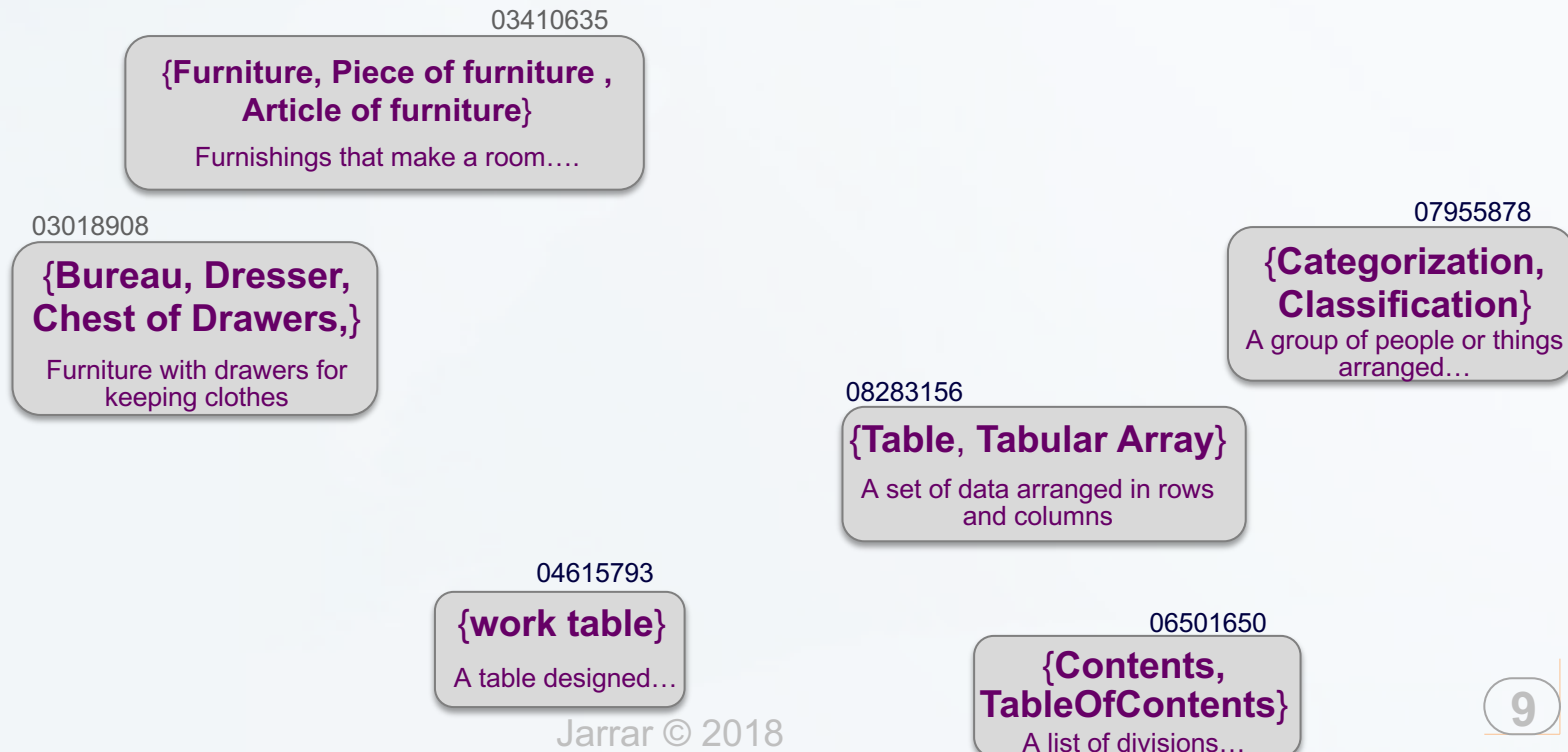
☐ Part 4: Global Wordnet

What is WordNet?

- In 1985 a group of **psychologists and linguists** at Princeton University started to develop a “**mental lexicon**”.
- You may also call it: “electronic dictionary”, “Mental dictionary”, English, “**semantic Network**”, hyperdimensional thesaurus, 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>

WordNet: Synonymy

- 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.
- Each synset is described by a **gloss** (examples of contexts).

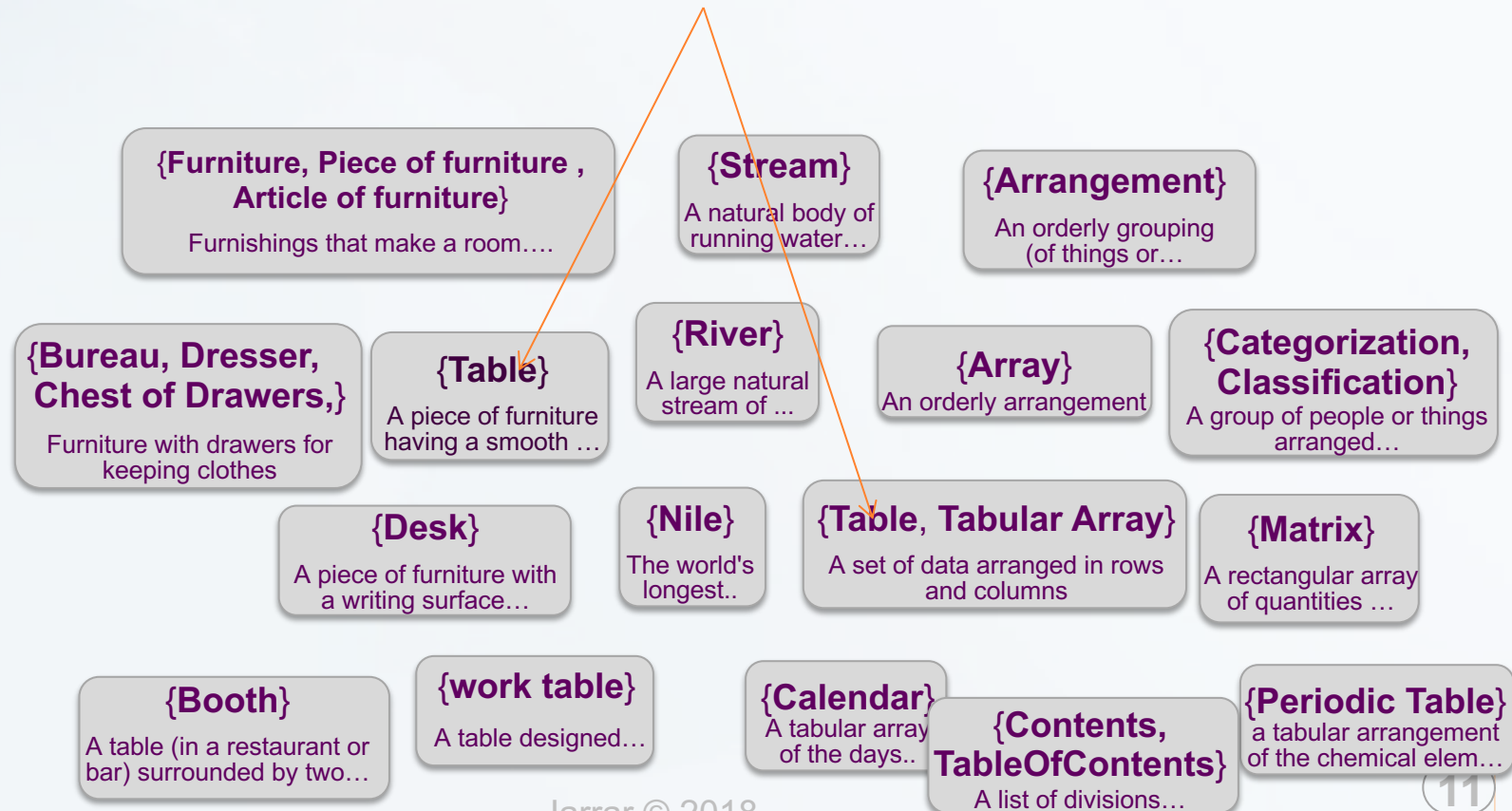


Exercise

List the different meanings of the words:
Table, Array, Matrix, Bureau

WordNet: Polysemy

- Each word form-meaning pair is unique.
- 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 synset.

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



WordNet: Statistics

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

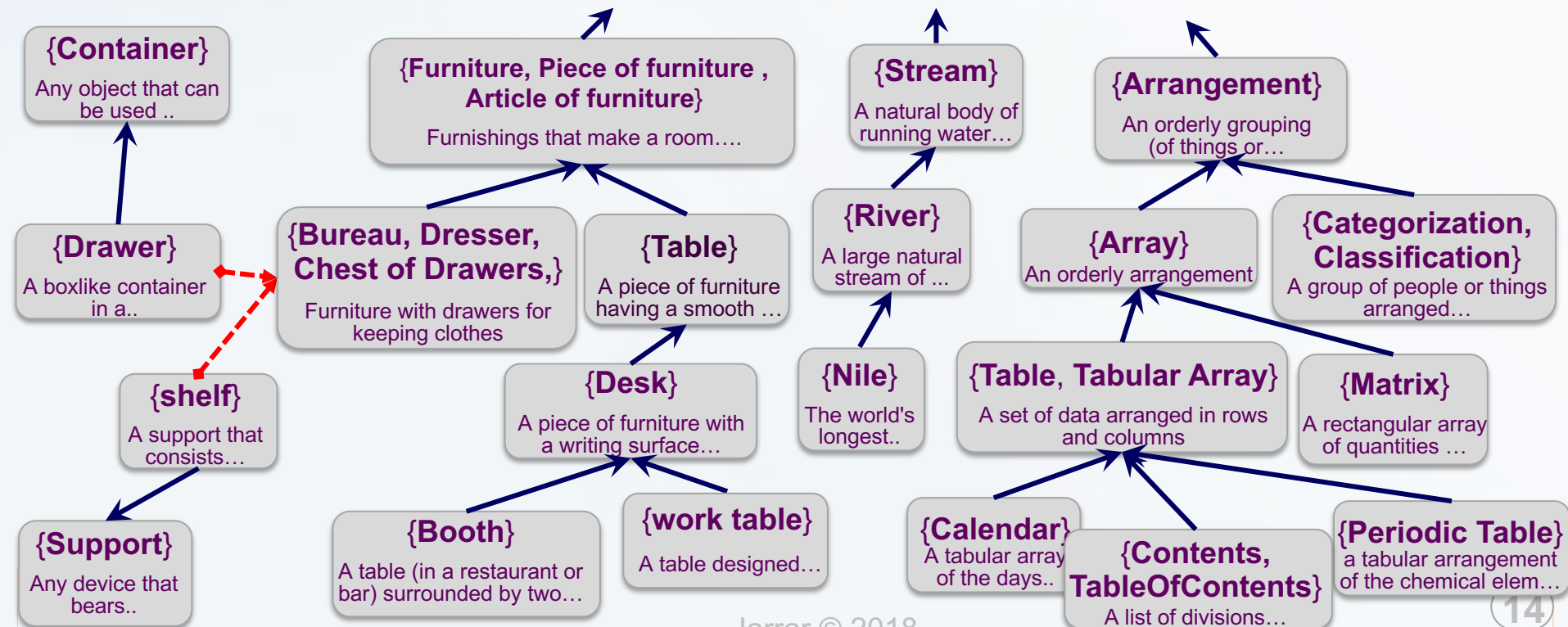


WordNet Semantic Relations

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

Such Relations are:

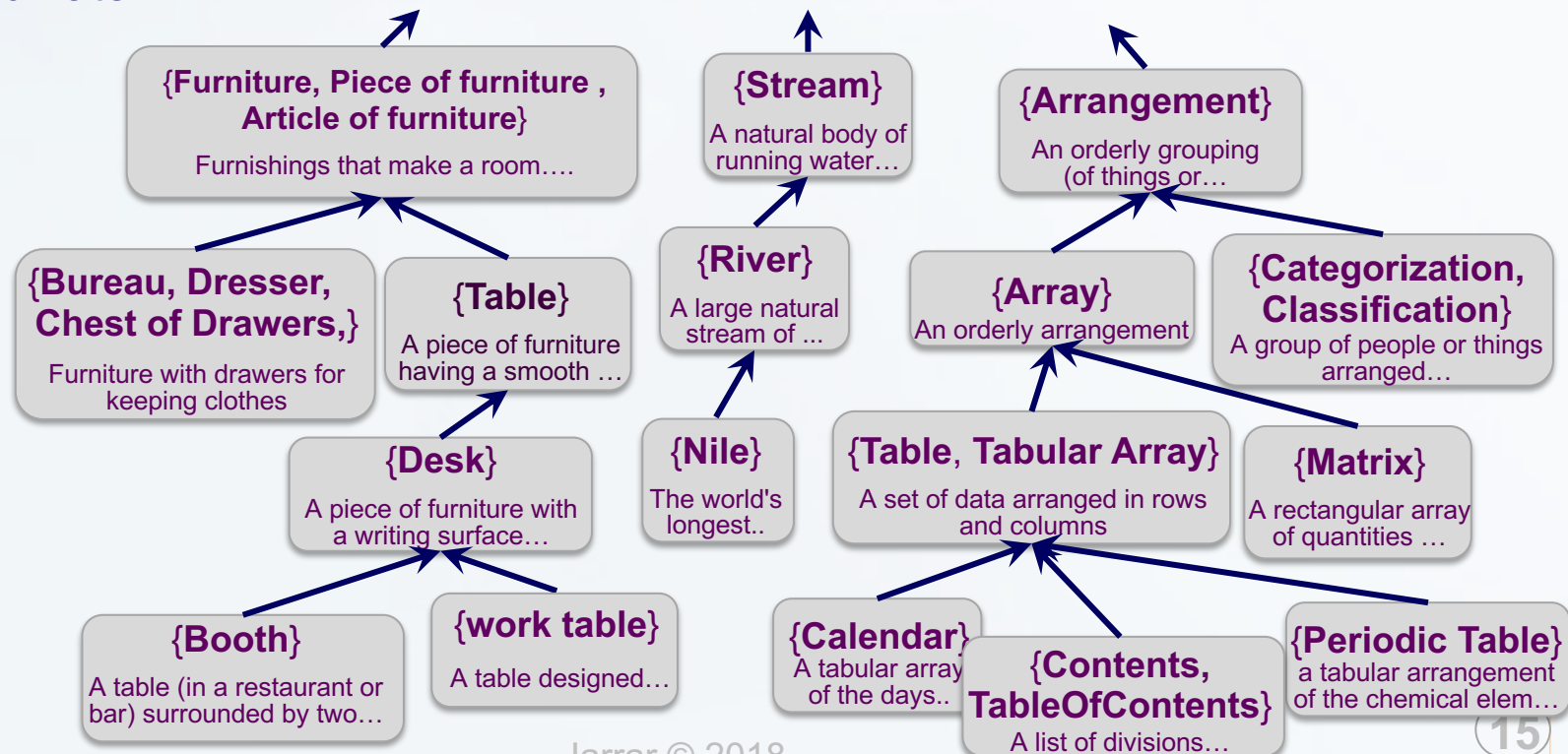
- **Hyponymy**, also called “Is a” relation, or sub/superordinate.
- **Meronymy**, also called “part of” relation



WordNet Relations: **Hyponymy**



- A synset $\{x, x', \dots\}$ is hyponym of the synset $\{y, y', \dots\}$ **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. [2]

Top Level Nouns (25 unique beginners)

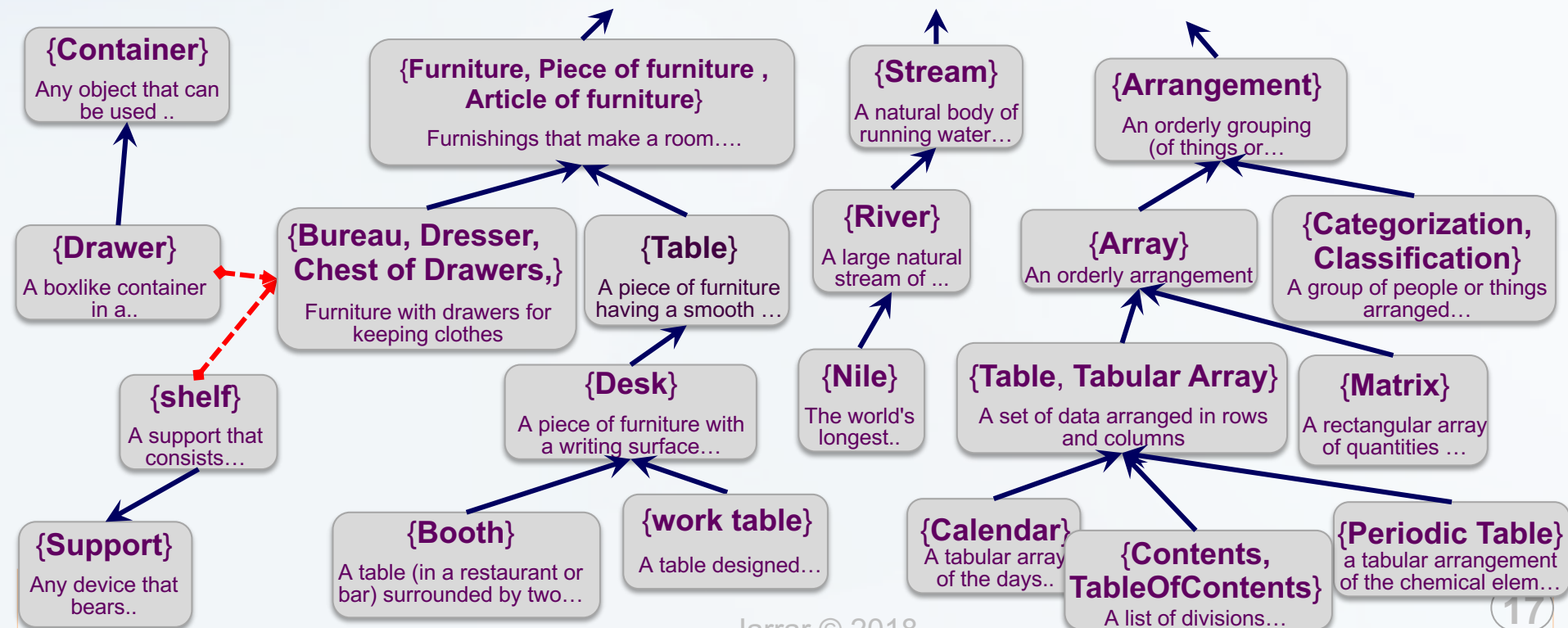
The WordNet hierarchy
is about **16 levels**

{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 $\{x, x', \dots\}$ is meronym of the synset $\{y, y', \dots\}$ 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

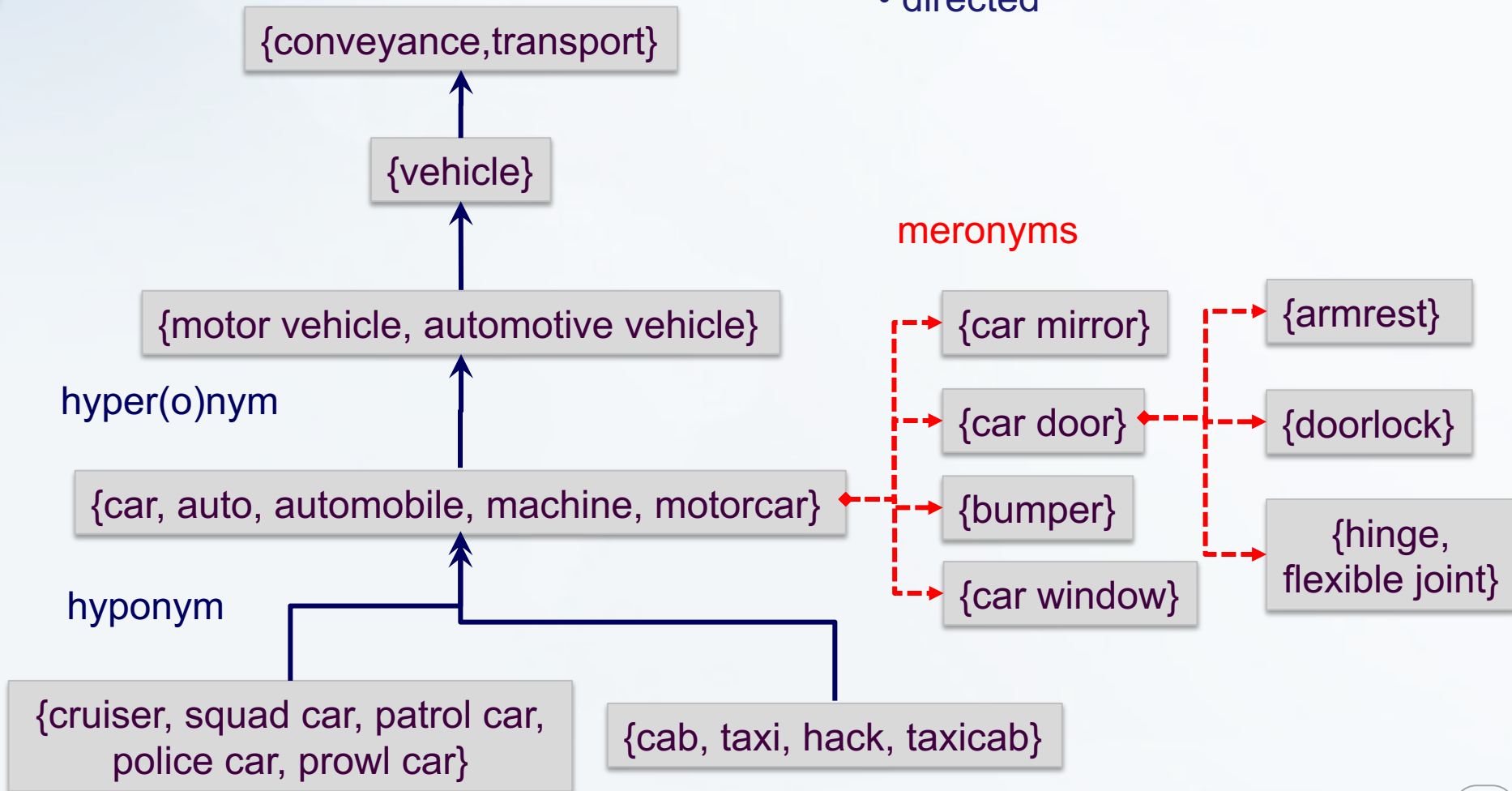
{car, auto, automobile, machine, motorcar}

WordNet Relations: Another Example

[Vossen]

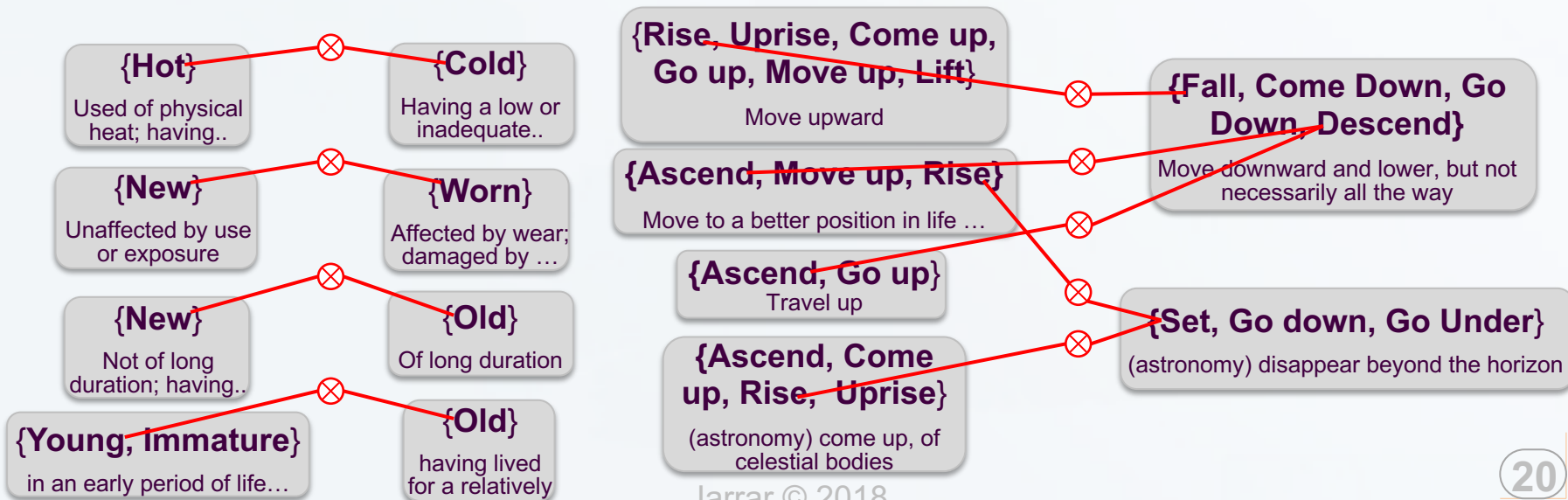
Hyponymy and meronymy relations are:

- transitive
- directed



WordNet Relations: Antonymy

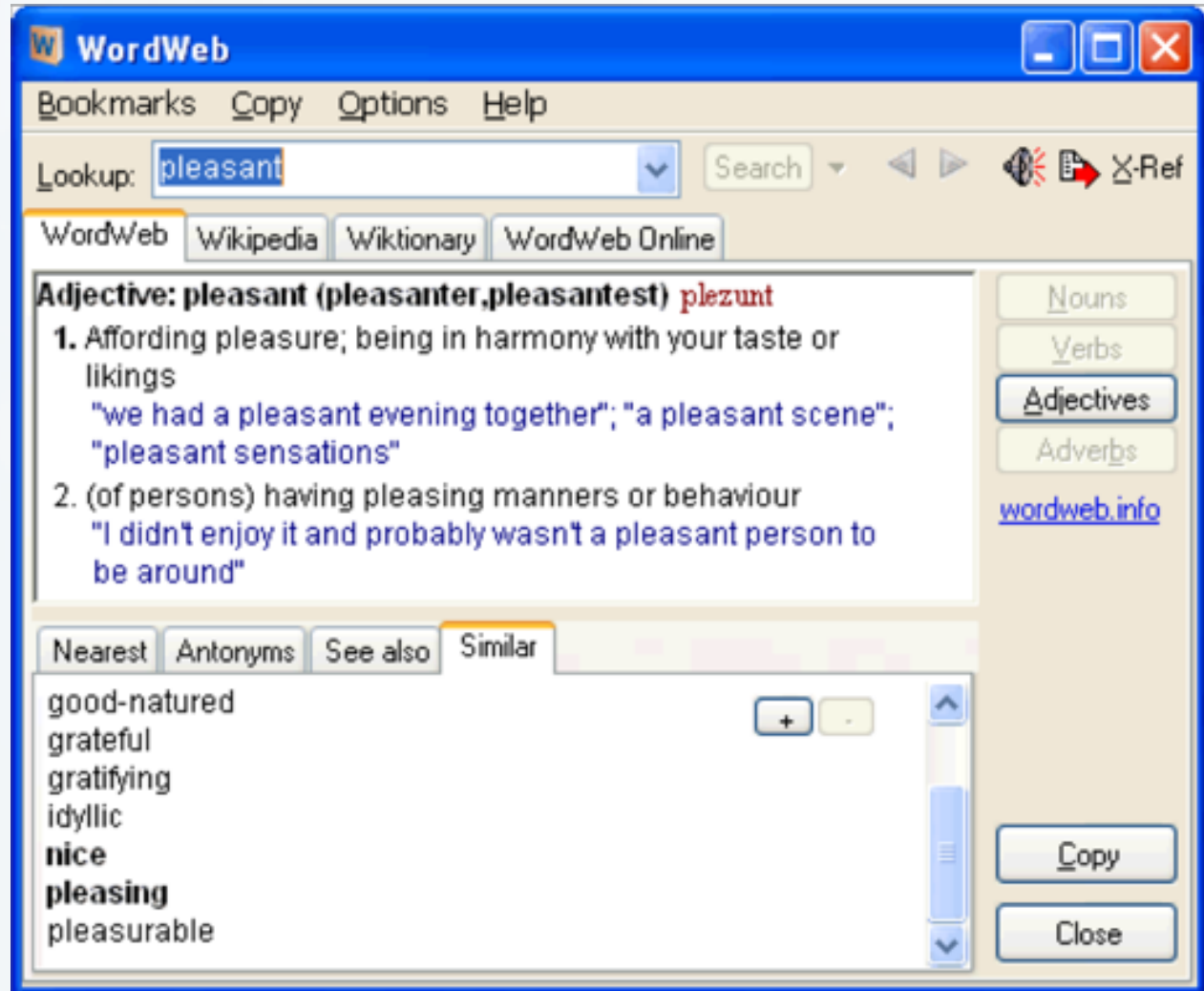
- The antonym of a word *x* is sometimes not-*x*, but not always. For example, *rich* and *poor* are antonyms, but to say that someone is not rich does not imply that they must be poor; many people consider themselves neither rich nor poor.
- Antonymy, which seems to be a simple symmetric relation, is 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*
- (Reflate₂ *Is-a* Inflate₃) (Inflate₃ *Is-a* Change₁) and (Reflate₂ *Is-a* Change₁) is **meaningless**, this is an implied relation.
- (Restrain₁ *Is-a* Inhibit₄) and (Inhibit₄ *Is-a* Restrain₁) 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.
- (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.

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

Introduction to Wordnets

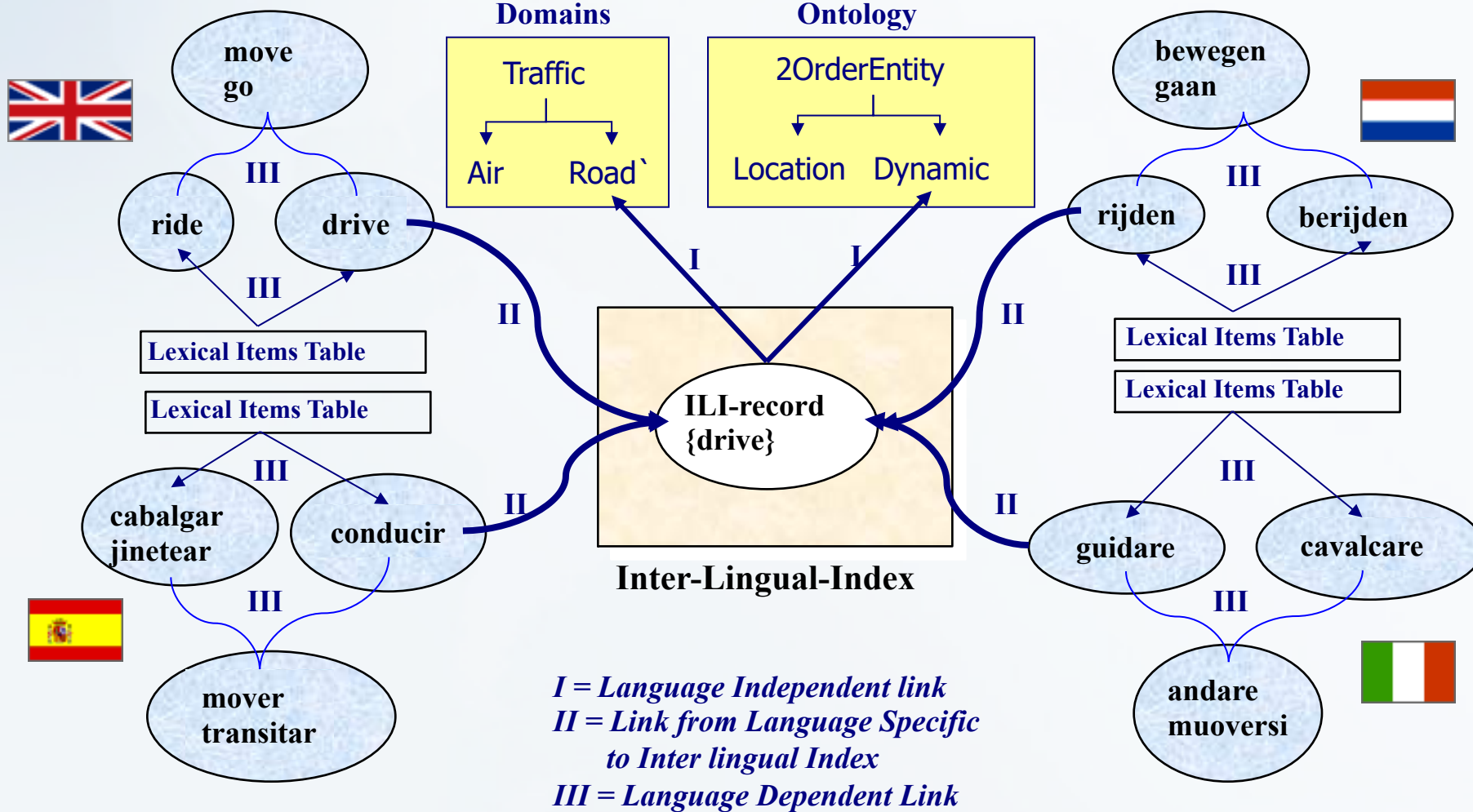
In this lecture:

- ☐ Part 1: What and why Thesaurus
- ☐ Part 2: What is WordNet
-  ☐ Part 3: **Euro Wordnet**
- ☐ Part 4: Global Wordnet

- 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;

- 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

Introduction to Wordnets

In this lecture:

- ☐ Part 1: What and why Thesaurus
- ☐ Part 2: What is WordNet
- ☐ Part 3: Euro Wordnet
-  ☐ Part 4: **Global Wordnet**

From EuroWordNet to Global WordNet

<http://www.globalwordnet.org>



The Global WordNet Association

- [GWC2012 Conference](#)
- [Wordnets in the world](#)
- [Wordnet Biblio](#)
- [Previous GWA Conferences](#)
- [Base Concepts](#)
- [The Global Wordnet Grid](#)
- [Membership form](#)
- [Mailing list](#)
- [The Constitution](#)
- [The Board](#)
- [Background document](#)

The Global WordNet Association is a free, public and non-commercial organization that provides a platform for discussing, sharing and connecting wordnets for all languages in the world. The aims of the association are:

1. To establish distribution facilities for the dissemination of the Association and Association publications and information materials:
 - To promote cooperation and information exchange among related professional and technical societies that build or use wordnets.
 - To provide information on wordnets to the general public.
2. To promote the standardization of the specification of wordnets for all languages in the world, including:
 - the standardization of the Inter-Lingual-Index for inter-linking the wordnets of different languages, as a universal index of meaning
 - the development of a common representation for wordnet data
3. To promote the development of sense-tagged corpora in all the linked languages.
4. To promote sharing and transferring of data, software and specifications across wordnet builders for different languages
5. To promote the development of guidelines and methodologies for building wordnets in new languages
6. To promote the development of explicit criteria and definitions for verifying the relations in any

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 50 languages, including:
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, will be explained later.

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.

References

[J17] Mustafa Jarrar: **Tutorial on Arabic Ontology Engineering**. The ACS/IEEE International Conference on Computer Systems and Applications. Tunis, 2017

1. Mustafa Jarrar, Hamzeh Amayreh: Linguistic Search Engine. Proceedings of the Web conference (WWW2019), ACM, 2019
2. Mustafa Jarrar: The Arabic Ontology - An Arabic Wordnet with Ontologically Clean Content. Applied Ontology Journal. IOS, 2019.
3. Mustafa Jarrar, Fadi Zaraket, Rami Asia, Hamzeh Amayreh: **Diacritic-Based Matching of Arabic Words**, ACM Transactions on Asian Language Information Processing. (Forthcoming)
4. Mustafa Jarrar, Werner Ceusters: **Classifying Processes and Basic Formal Ontology**, Proceedings of the 8th International Conference on Biomedical Ontology (ICBO 2017), Newcastle, UK. 2017
5. Mustafa Jarrar: **Building a Formal Arabic Ontology (Invited Paper)**. In proceedings of the Experts Meeting on Arabic Ontologies and Semantic Networks. Alecco, Arab League. Tunis, July 26-28, 2011.
6. 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.
7. Mustafa Jarrar, Anton Deik: **The Graph Signature: A Scalable Query Optimization Index for RDF Graph Databases Using Bisimulation and Trace Equivalence Summarization**, International Journal on Semantic Web and Information Systems, 11(2), 36-65, DOI: 10.4018/IJSWIS.2015040102. April-June 2015
8. Mustafa Jarrar: Lecture Notes on Lexical Semantics and Multilingualism. Birzeit University, Palestine. 2018
9. Mustafa Jarrar: Lecture Notes on Introduction to Wordnets. Birzeit University, Palestine. 2018
10. Mustafa Jarrar: Lecture Notes on the Arabic Ontology Basics. Birzeit University, Palestine. 2018
11. 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). ISBN 978-88-903120-2-1. Campione, Italy. June, 2011.
12. 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).
13. 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. Portugal. November, 2007
14. Mustafa Jarrar, Maria Keet, and Paolo Dongilli: **Multilingual verbalization of ORM conceptual models and axiomatized ontologies**. Technical report. STARLab, Vrije Universiteit Brussel, February 2006.

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

<http://iaoa.org/isc2012/docs/fellbaum-trento-uno.pdf>

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

[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), 288-298.

[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, ISBN 0415131596.

[D92] Davidson, Herbert Alan (1992), *Alfarabi, Avicenna, and Averroes on Intellect: Their Cosmologies, Theories of the Active Intellect, and Theories of Human Intellect*, Oxford University Press, p. 146, ISBN 0195074238