Announcements

Office hour: Wednesday July 20 11-12

Cubicle No. 1 IBS

You can also e-mail me (fellbaum@princeton.edu)with questions or make a personal appointment

Happy Hour for Computational linguists July 18 5:30

Dark Horse bar on Baseline Rd. between 29th & 30th streets, near William's Village

Special lecture on the Sketch Engine

Adam Kilgarriff Wednesday July 20 4:30-5:30 PM

145 Eaton Humanities

WordNet
(Part Two)

Christiane Fellbaum

Synonymy, polysemy

WordNet gives information about two fundamental, universal properties of human language:

synonymy and **polysemy**

Synonymy = one:many mapping of meaning and form

Polysemy = one:many mapping of form and meaning

Basic relation: synonymy

Each node in the semantic network is a “concept”
“Concept” is expressed by several different word forms
Synonym sets (“synsets”) are the building blocks of WordNet

{beat, hit, strike}
{car, motorcar, auto, automobile}
{ big, large}
{queue, line}

Synset members are unordered
All express/denote/refer to the same concept
WN disregards differences in frequency, connotation, register, genre…
“cognitive synonymy” (Cruse 1986)

Polysemy

One word form expresses multiple meanings

{*table*, tabular_array}
{*table*, piece_of_furniture}
{*table*, mesa}
{*table*, postpone}

Polysemy in WordNet

A word form that appears in *n* synsets
is *n*-fold polysemous

{*table*, tabular_array}
{*table*, piece_of_furniture}
{*table*, mesa}
{*table*, postpone}

table is fourfold polysemous/has four senses

Some WordNet stats

Part of speech	Word forms	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

Note that WordNet in fact consists of four distinct networks, one for each POS

Few connections across synsets with words from different POS

WordNet stats

The figures include

--some phrases with questionable lexical status ("change integrity"), often needed to label a category and distinguish it from others (more later)

--a somewhat random selection of proper names (people, places, products); many more could/should be added

The “Net” part of WordNet

Synsets are interconnected

Bi-directional arcs express semantic relations

Result: large semantic network

(directed acyclic graph/DAG)

Whence the relations?

Classical ontology (Aristotle):

IS-A (kind/type of): poodle-dog

HAS-A (part): dog-tail

Co-occurrence patterns in texts:

meaningfully related words are used together

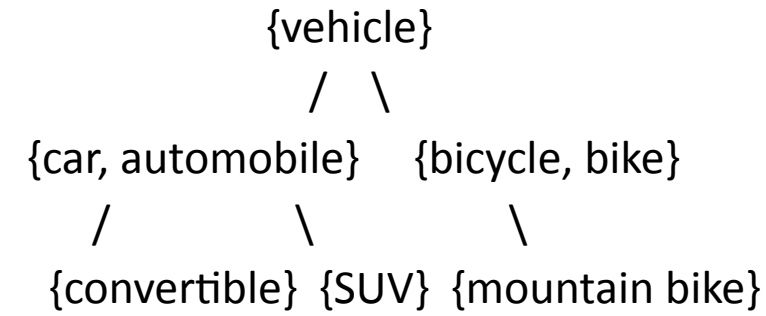
or show similar distribution (more on that later)

Word Association norms

Hypo-/hypernymy relates noun synsets

Relates more/less general concepts

Creates hierarchies, or “trees”



“A car is is a kind of vehicle” <=>“The class of vehicles includes cars, bikes”

Noun hierarchies can have up to 16 levels

Tree(s)

About a dozen high-level concepts:

person, animal, artifact, location, motion, communication,…

All link to a single root, *entity*

(This allows programs to compute the distance between ANY two nodes)

Hyponymy

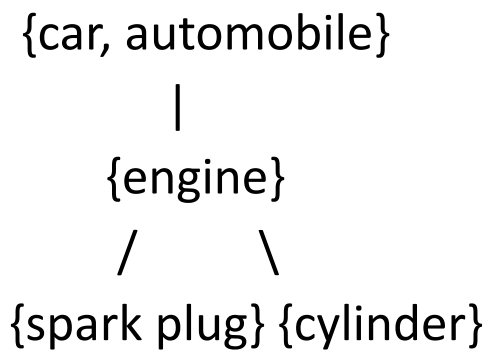
Transitivity:

A car is a kind of vehicle

An SUV is a kind of car

=> An SUV is a kind of vehicle

Meronymy/holonymy (part-whole relation)



“An engine has spark plugs”

“Spark plus and cylinders are parts of an engine”

Meronymy/Holonymy

Inheritance:

A finger is part of a hand

A hand is part of an arm

An arm is part of a body

=>a finger is part of a body

(Note that statements like “a fingernail is a part of an arm” seem odd--though they are true--while others like “a fingernail is a part of the body” seem natural. Why is that?)

Meronymy

WordNet distinguishes three kinds of meronymy

Proper parts (count nouns):

arm-body, page-book, *branch-tree*

Substance/Stuff (mass nouns):

oxygen-water, flour-pizza

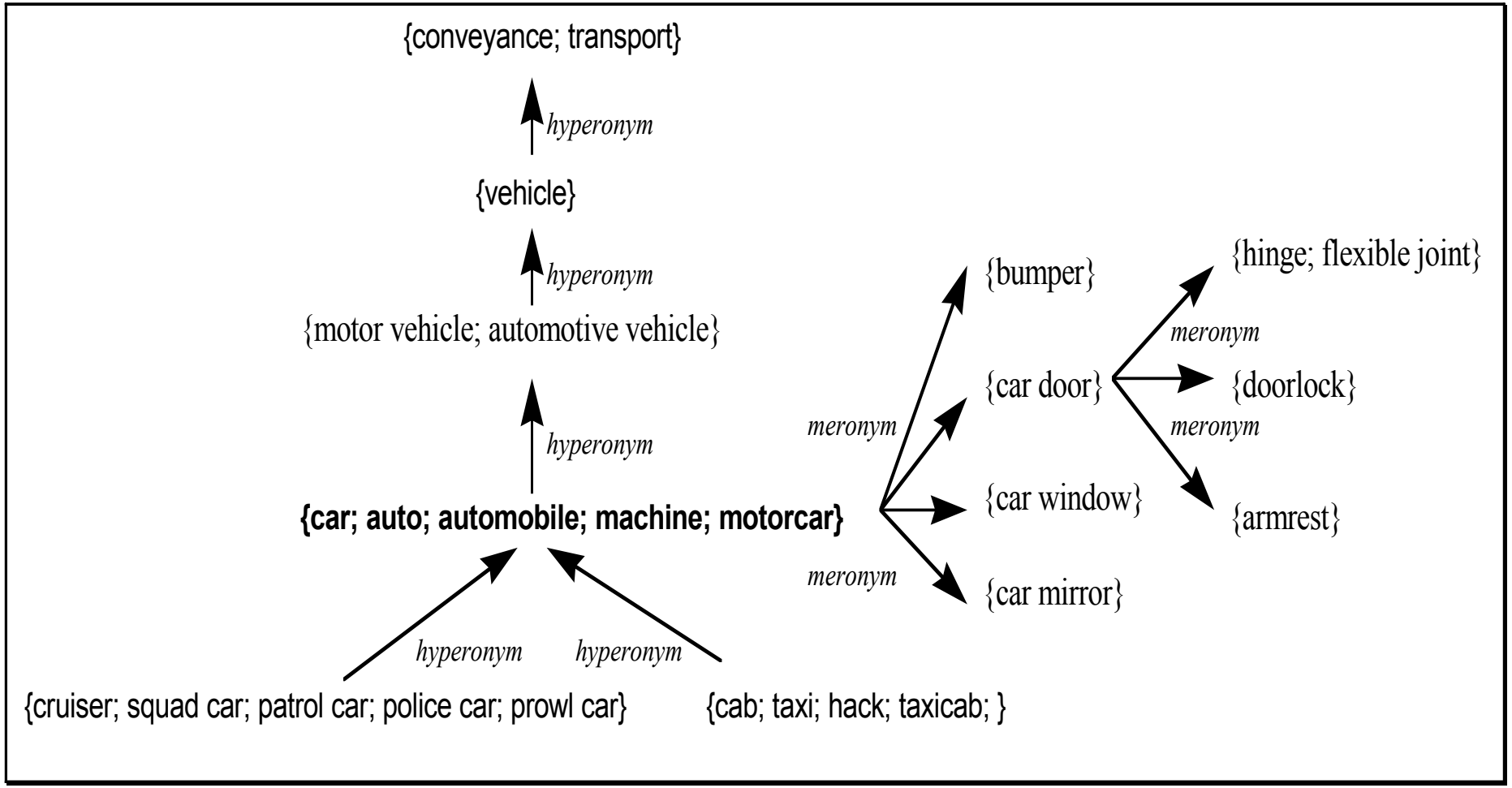
Member-group*:*

student-class, tree-forest, bird-flock

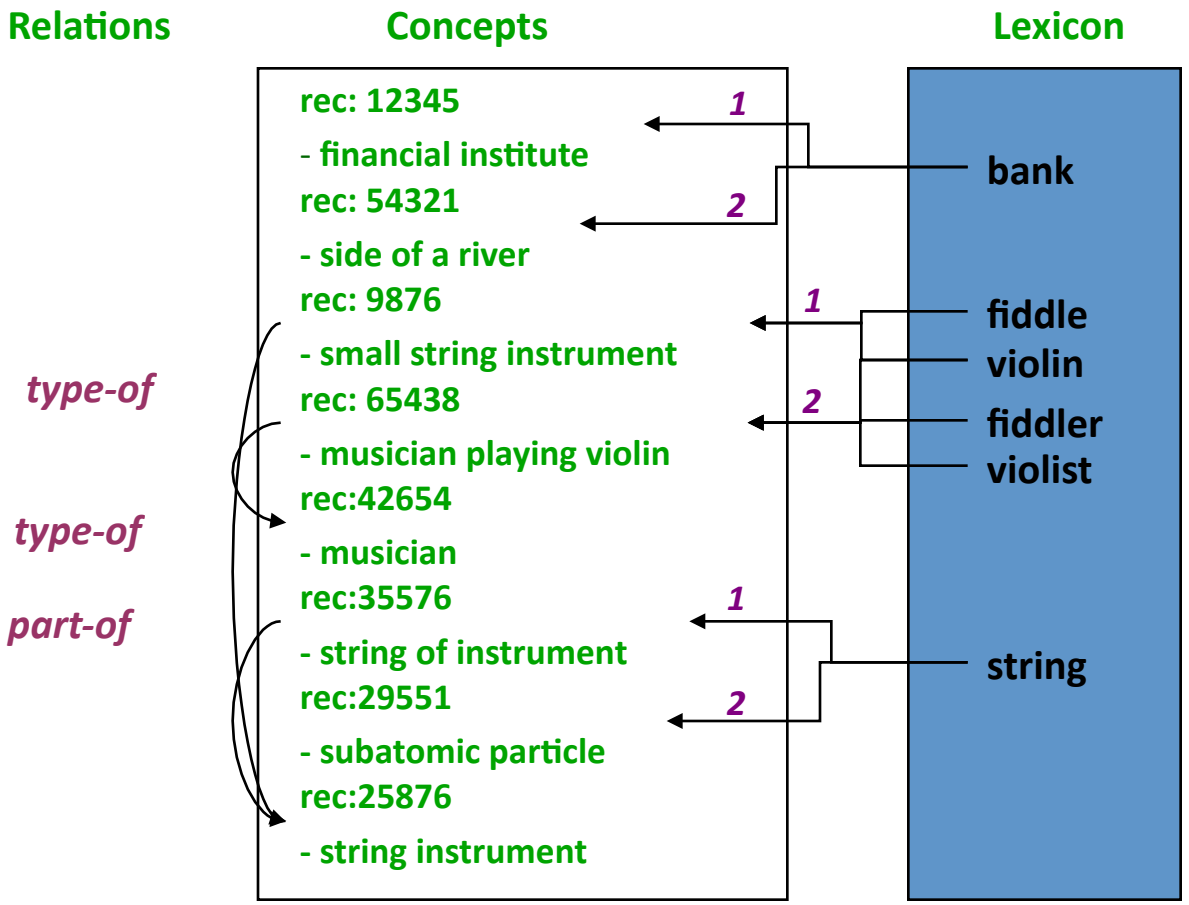
(the whole would not exist but for the members)

There are arguably more kinds of meronymy

Structure of WordNet (Nouns)



WordNet Data Model



Adjective relations: antonymy

Strong mutual association between members of antonymous adjective pairs:

hot-cold, old-new, high-low, big-small,...

Distributional overlap (shared selectional restrictions)

Highly frequent, polysemous

Statistically high co-occurrence in the same sentence (Justeson and Katz 1991)

Members of antonymous pairs are acquired together by children

Adjective relations

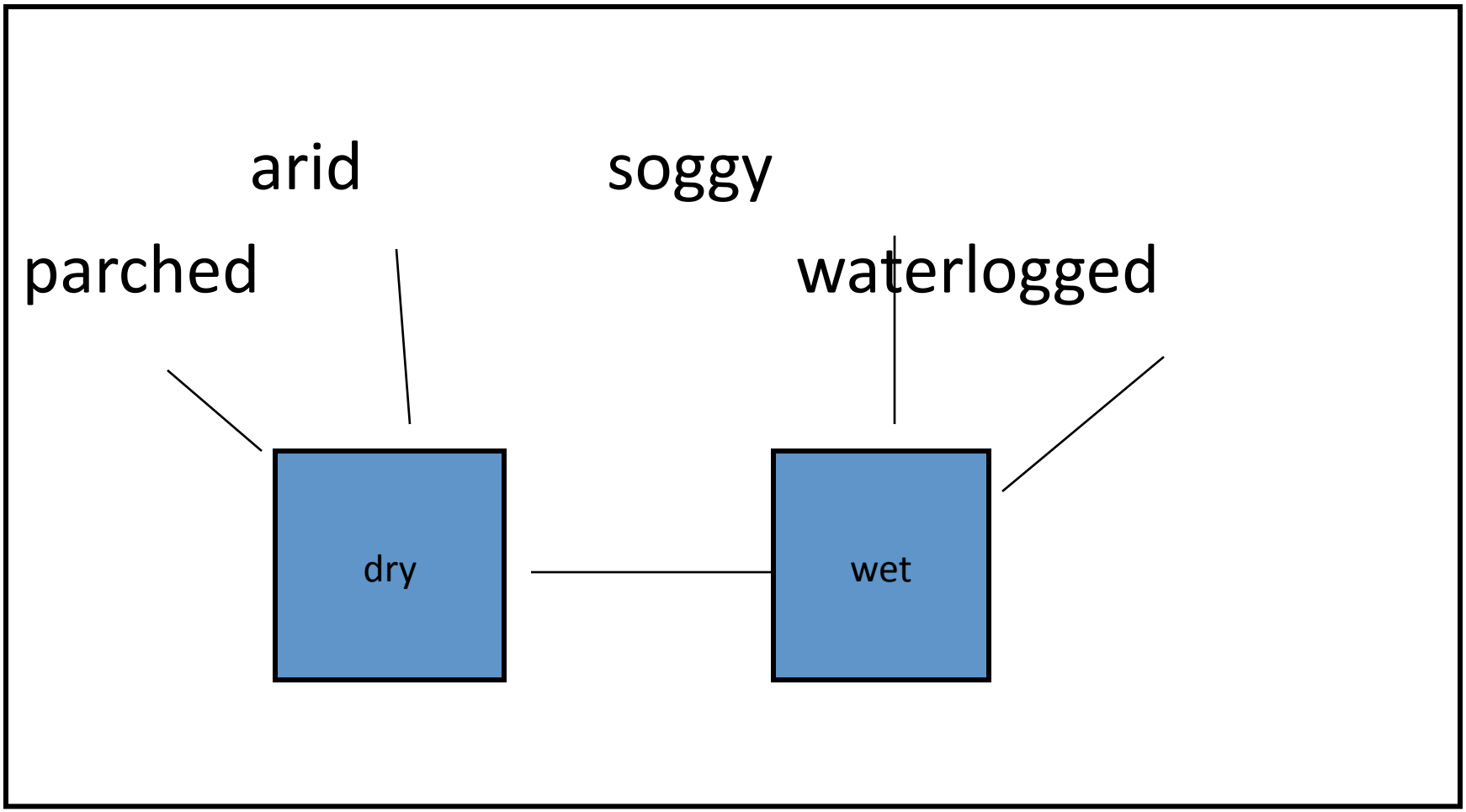
WordNet connects members of pairs like

hot-cold, long-short, new-old, wide-narrow,…

(“direct antonyms”)

For each adjective may there may be similar but less salient ones (e.g., *cool, lengthy, ancient,...)*

The “dumbbell” model



“Dumbbell” model

- Direct antonyms: *dry-wet, long-short, old-new, high-low,* etc.
- Indirect antonyms are “similar” to one member of the “dumbbell”

Experimental evidence

Reaction time for responses to questions like

"Is *dry* the opposite of *wet*?" (direct antonyms)

"Is *dry* the opposite of *waterlogged*?"

(direct-indirect)

"Is *arid* the opposite of *waterlogged*?"

(indirect-indirect)

Gross, Fischer, Miller (1989)

Experimental evidence

- Fastest response: direct-direct pairs
- Less fast: direct-indirect pairs
- Hesitation/slow response: indirect-indirect pairs

Problem: word frequency is strongly correlated with response time. More frequent words are accessed faster than rare word.

remainders

Not all adjectives fit into dumbbells

"Pertainyms" are derived from and linked in WordNet to nouns (*political-politics, nuclear-nucleus,* etc.)

Problem: adjectives that have no antonyms (*angry*)

Relations among verbs

Manner relation ("troponymy")

to x is to y in some manner

connects verbs like

move-walk, whisper-talk, smack-hit, gobble-eat

Can construct trees (not as deep as nouns):

move-run-jog-run

communicate-talk-whisper

Troponymy is polysemous: specific manner depends on verb category

Verb trees

No single top node: hundreds of flat "bushes" with no more than 5 levels

(what would a top node be?)

High-level nodes:

Verbs of motion, change of state, communication, cognition, contact, consumption, etc.

Other relations among verbs reflect temporal or logical order between two events

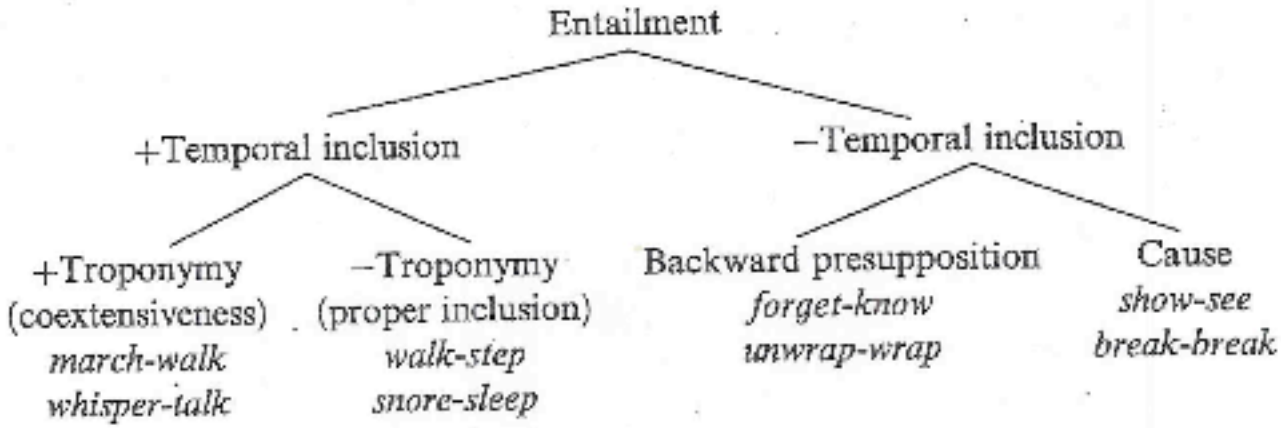
divorce-marry (backward presupposition)

snore-sleep, pay-buy (inclusion)

kill-die, fell-fall (cause)

One event unidirectionally entails the other

Entailment also holds among troponyms


Entailment
+Temporal inclusion
−Temporal inclusion
+Troponymy
(coextensiveness)
march-walk
whisper-talk
−Troponymy
(proper inclusion)
walk-step
snore-sleep
Backward presupposition
forget-know
unwrap-wrap
Cause
show-see
break-break

Questionable entries

{change_integrity} has subordinates *break, shatter, explode, evaporate, liquify*,….

{change_shape} has subordinates *bend, roll_up, twist,…*

Similar: *change_magnitude, change_location,…*

Lexically odd but these “words” allos cleaner semantic (and syntactic) distinction among large categories (trees)

Syntactically motivated "polysemy"

Example: spray-load alternation motivated separate entries for verb variants

(1) Superordinate synset {*distribute, put*}

has subordinate synsets *spray, spritz, squirt*,…

as in

Spray paint onto the canvas

Squirt water on the counter

Material (locatum)-location

(2) Another superordinate synset {cover} has same subordinate verbs, but with different syntax:

Spray the canvas with paint

Sprinkle the lawn with water

Location-locatum (material)

So: notion of synset is broad (not strictly equivalent to “sense”)

Many synsets represent syntactically distinct variants

Others serve as broad categories with large number of members

WN as a lexical resource

"Have concept, need words"

depart from synset, travel in WordNet space

"Have word, need concept"

query word, find matching synsets

Is WordNet a Thesaurus?

Kind of:

groups together meaningfully related words

No:

limited number of explicitly labeled relations

related words are linked to specific concepts (disambiguated); thesaurus is a “bag of words”

many words linked in WordNet do not co-occur in the same thesaurus entry

WordNet allows one to measure and quantify the semantic similarity or distance among words and concepts (more in next lecture)

Constructing new wordnets
Augmenting wordnet

Princeton WordNet (English) was manually built

Many new wordnets don't have resources for manual construction

Bootstrap it!

Use bootstrapping methods for augmenting existing WordNet

Semi-automatic WN construction/ enrichment

Lexical-semantic patterns (Cruse 1986)

Pattern for super-subordinate relations:

Xs and other Ys (*roses and other flowers*)

Ys such as Xs (*flowers such as roses*)

Can search a corpus with patterns to identify general-specific pairs

Augment existing WordNet with domain-specific terms, proper names

(Hearst 1993, Snow et al. 2004)

Lexical-semantic patterns

Given known pairs in a specific relation, induce additional patterns from a corpus/Web, using a wildcard

rose * flower

flower * rose

Other patterns/for other relations (meronymy, verb relations)?

(of course you'll get some noisy patterns and noisy data when using the patterns)

Lexical-Semantic Patterns

Some adjectives have scalar properties

?comfortable<prosperous<rich<well-heeled<affluent<wealthy

Scalar representation would be superior to dumbbells with semantically undifferentiated “similar” adjectives

Intuitions about scalar ordering are not always clear

Corpus data provide support

AdjScales (work in progress)

Induce patterns using centroid and similar from WordNet

*well-off * rich*

*destitute * poor*

Bootstrap and apply patterns

(Sheinman and Tokunaga 2009, Sheinman et al. 2011)

AdjScales (work in progress)

Two kinds of patterns

"strong" : more intense verb on the right

poor if/but not destitute

poor (perhaps) even destitute

"weak" : less intense verb on the right

if not destitute, (at least) poor

not destitute but still poor

(Sheinman & Tokunaga 2009; Sheinman et al. 2011)

Crosslinguistic WordNets

Starting in late 1990s, WordNets were built for languages other than English

Genetically and typologically unrelated languages: Turkish, Hindi, Chinese, Korean, Basque, Xhosa, Arabic, Latin… (currently >60)

All are mapped to Princeton WordNet

Great potential for crosslinguistic applications

http://www.globalwordnet.org

Crosslinguistic WordNets

Some are manually constructed

--independently from PWN, mapped later

or

--translated directly from PWN

First method is considered easier, more accurate (why?)

Crosslinguistic WordNets

Semi-automatic methods:

--apply lexical-semantic patterns to corpora

--parse dictionary definitions that employ patterns, e.g.

an X is a Y that…. => *Y* is a superordinate of *X*

(do patterns hold crosslinguistically?)

--parallel (aligned) corpora

--enlist the anonymous masses (Amazon Mechanical Turk)

Mapping words and synsets across multilingual WordNets

First set of foreign-language WNs were built with reference to Princeton WordNet

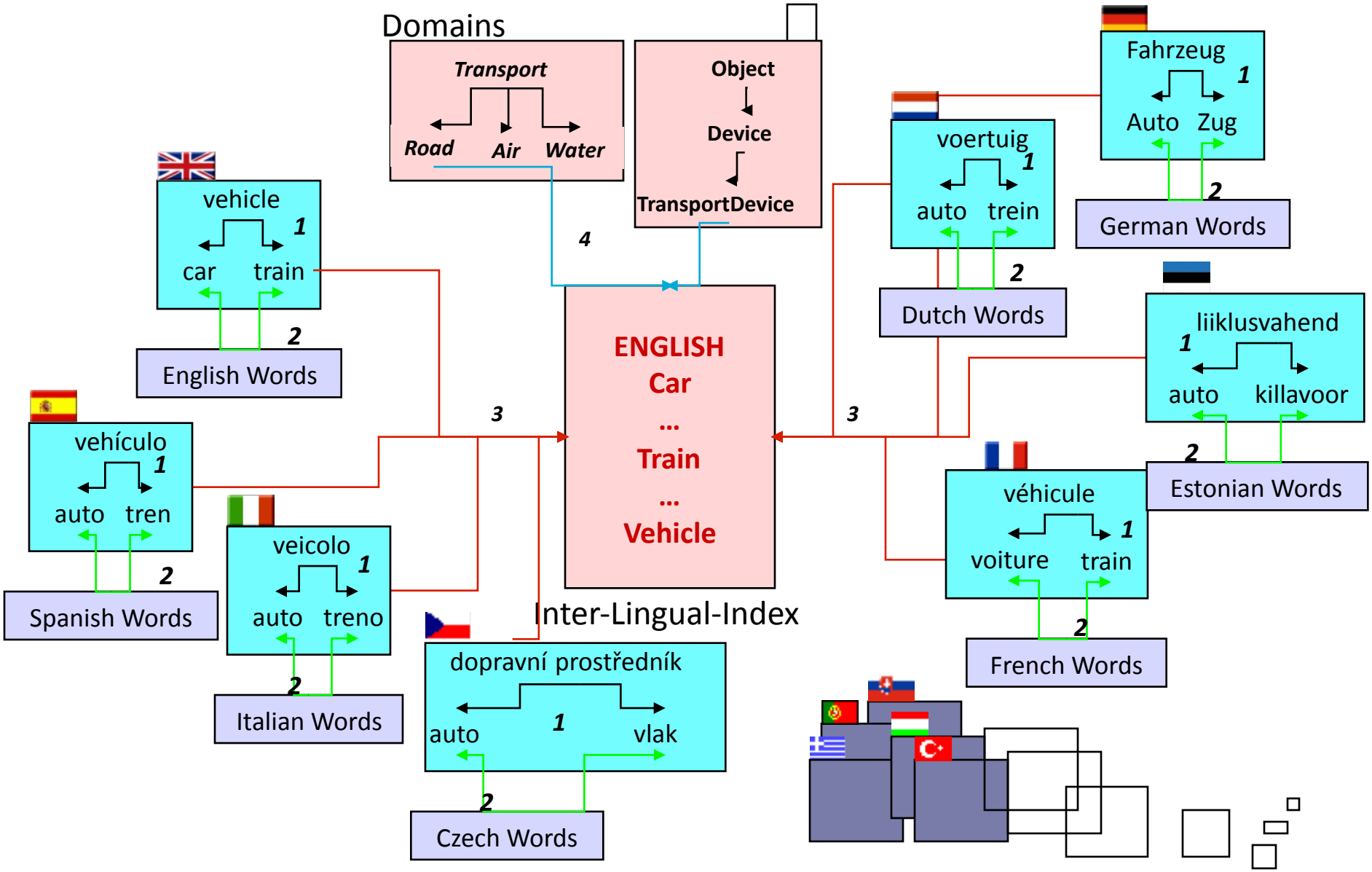
Princeton WN as the hub ("interlingual index")

Each synset in each WN was linked to a "record" (PWN synset identifier) in the index

Crosslingual mapping of words and synsets proceeds via the index

The index is a flat, unstructured list

Only language-specific wordnets have relations and form networks


Domains
Transport
Road
Air
Water
Object
Device
TransportDevice
4
ENGLISH
Car
...
Train
...
Vehicle
Inter-Lingual-Index
3
3
vehicle
1
car
train
2
English Words
vehículo
1
auto
tren
2
Spanish Words
veicolo
1
auto
treno
2
Italian Words
dopravní prostředník
auto
1
vlak
2
Czech Words
voertuig
1
auto
trein
2
Dutch Words
Fahrzeug
1
Auto
Zug
2
German Words
liiklusvahend
1
auto
killavoor
2
Estonian Words
véhicule
1
voiture
train
2
French Words

Mismatches in multilingual WordNets

Concepts not lexicalized in English required new records (w/out English synset):

--Arabic lexically distinguishes 12 kinds of *cousin*

Conversely, some lgs. lack equivalents of English words:

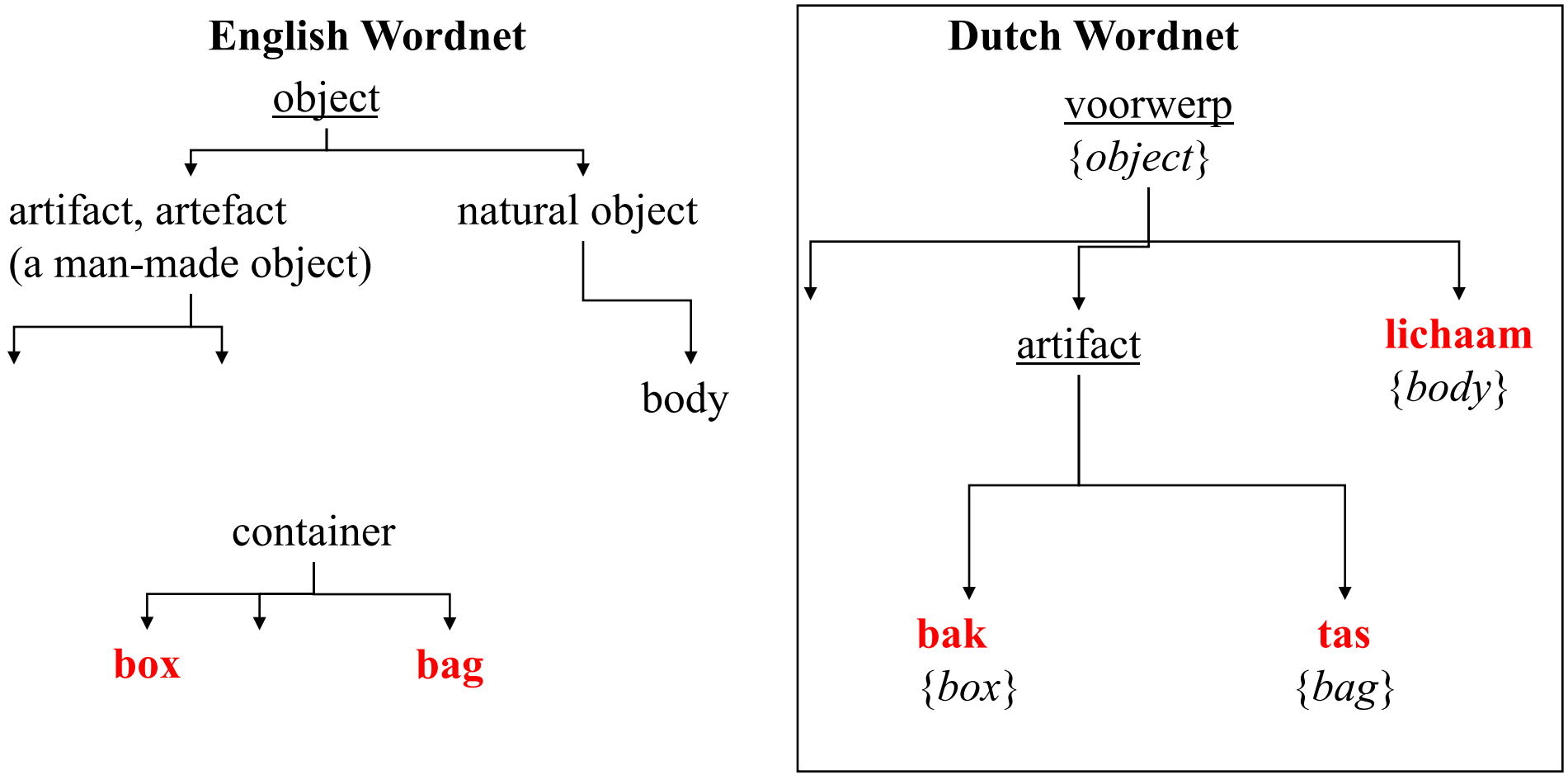
--Dutch lacks *container* but lexicalizes kinds (hyponyms) of *container* (*box, bag, bottle,..*)

Respective hierarchies reflects this difference:

Du. *bag, box,..=>artifact*

Engl. *bag, box,…=>container =>artifact*

English-Dutch snippet



What is universal?

- Surely not all “concepts”:
- English has many verbs of walking (*slouch, strut, stroll,amble, prance, sneak, march*,..) and walking/running (*hop, skip, bounce*,…)
- No 1:1 crosslingual encoding of concepts
- But is the network structure universal?

Multilingual WordNets

Interlingual Index is biased towards English

Could skew coverage of new wordnets that are translated from English

Can't always map across languages

Solution: replace index by *language-independent, formal* ontology

Meanings are stated as axioms in logical form

Automatic systems can process entries

Enables automatic reasoning, crosslinguistic applications