

Worlds of words

Piek Vossen
Faculty of Arts



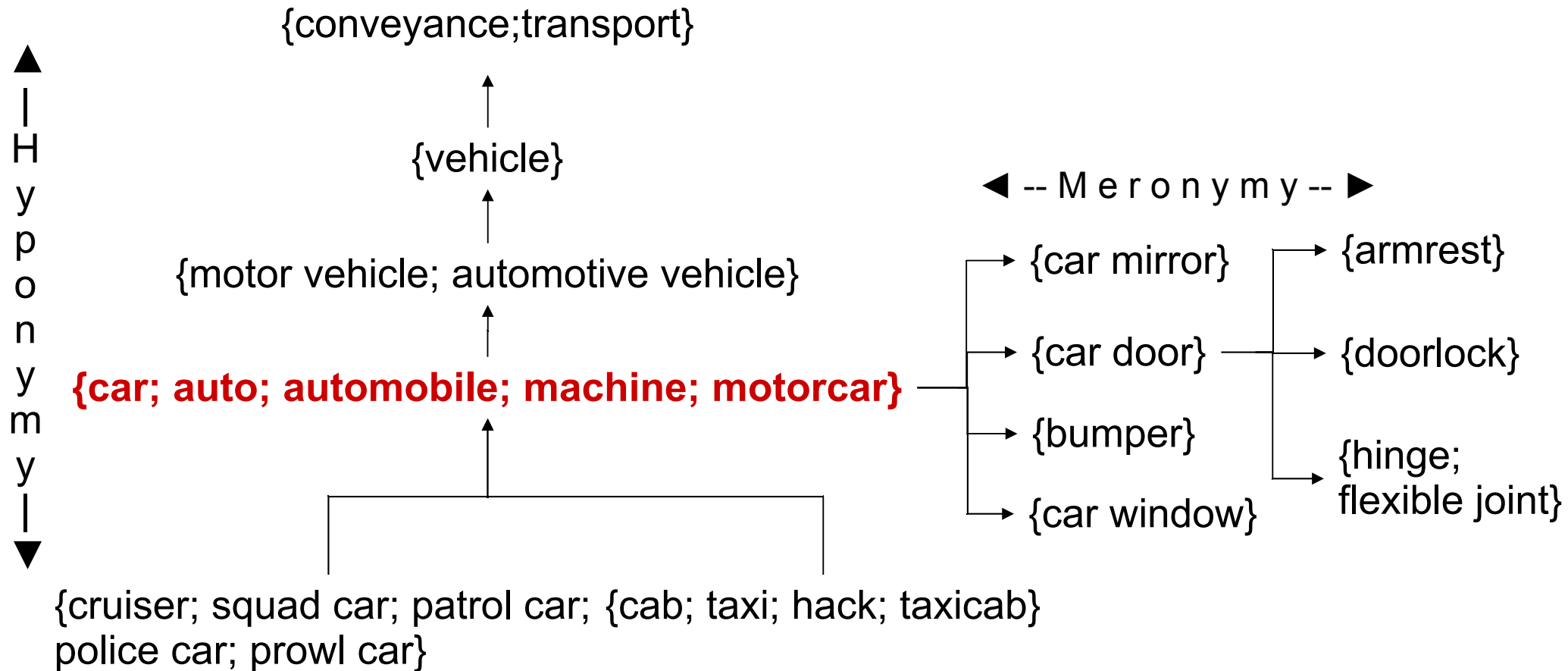
Worlds of words

- Small world networks tend to be more efficient through hubs and robust as long as hubs are not attacked
- The words of languages can be viewed as networks in many different ways:
 - Words connected to each other on the basis of their **form**: *paard, baard, waard, kaart, vaart, gaart, maart* (see Jeronimus, Westerveld, van den Berg and van Leeuwen 2009 for Japanese Kanji)
 - Words connected by their **co-occurrence in text**
 - Words connected by their **meaning**

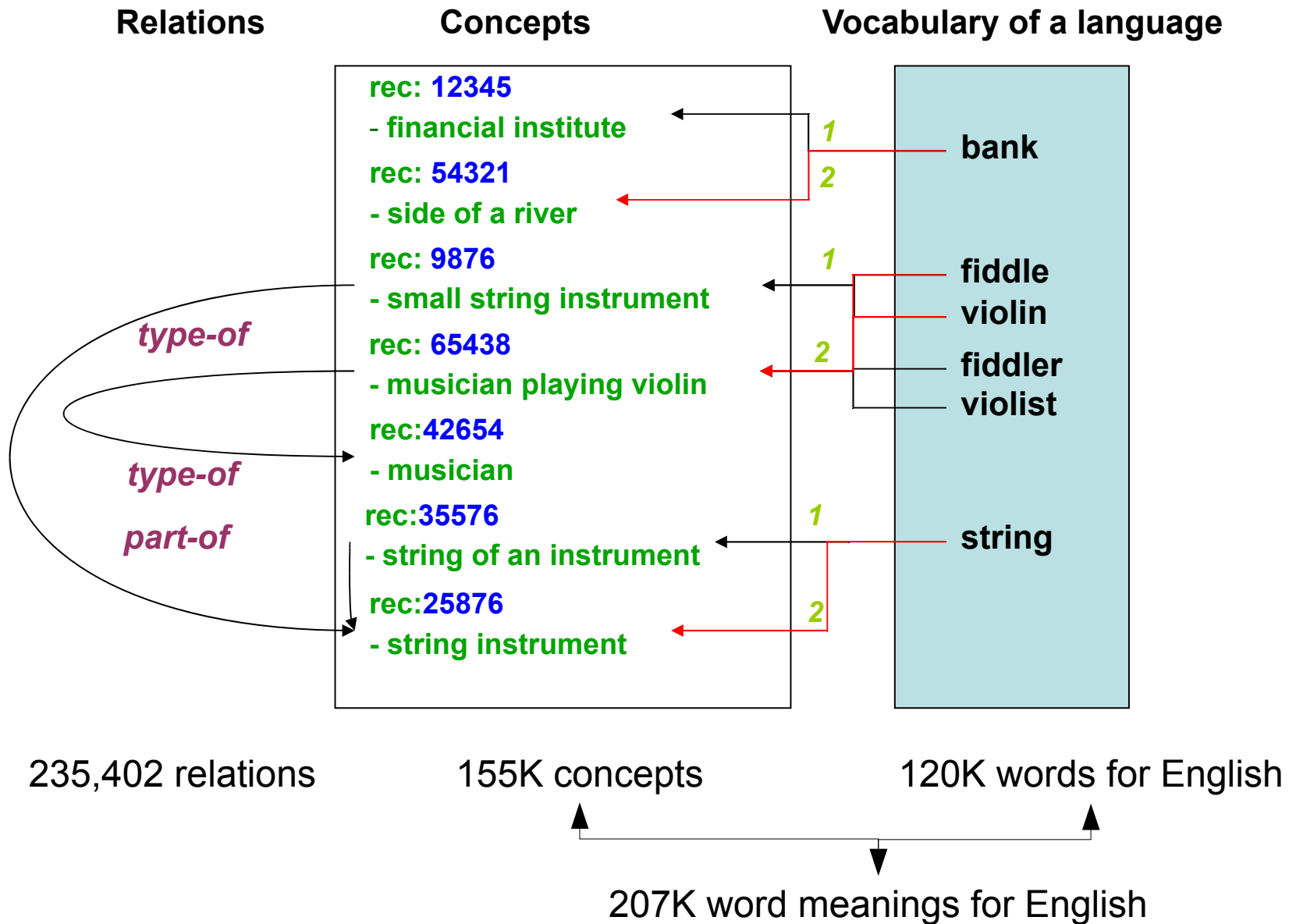
Word networks of co-occurrence

- Study by Ferrer I Cancho & Solé
- Words are the vertices and their context words in text constitute relations:
 - “look at the size of the **head** of the **crocodile** and see its tail also protruding from the water”
 - “Walking the long haul to the **head** of the **traffic jam**”
 - “The **head** of the **department** will report to the directorate”
- “head” will have many connections e.g. to “department”, “traffic jam” and “crocodile”
- Co-occurrence patterns represent important semantic structures which show small world properties.
- The most frequent words with most meanings (e.g. head has 30 meanings) are the hubs!!

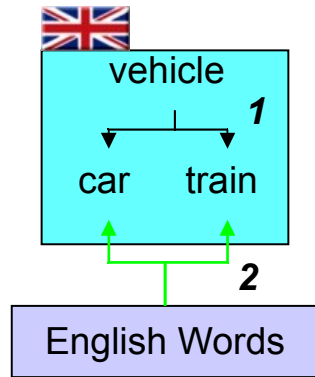
Wordnet: a network of word meanings



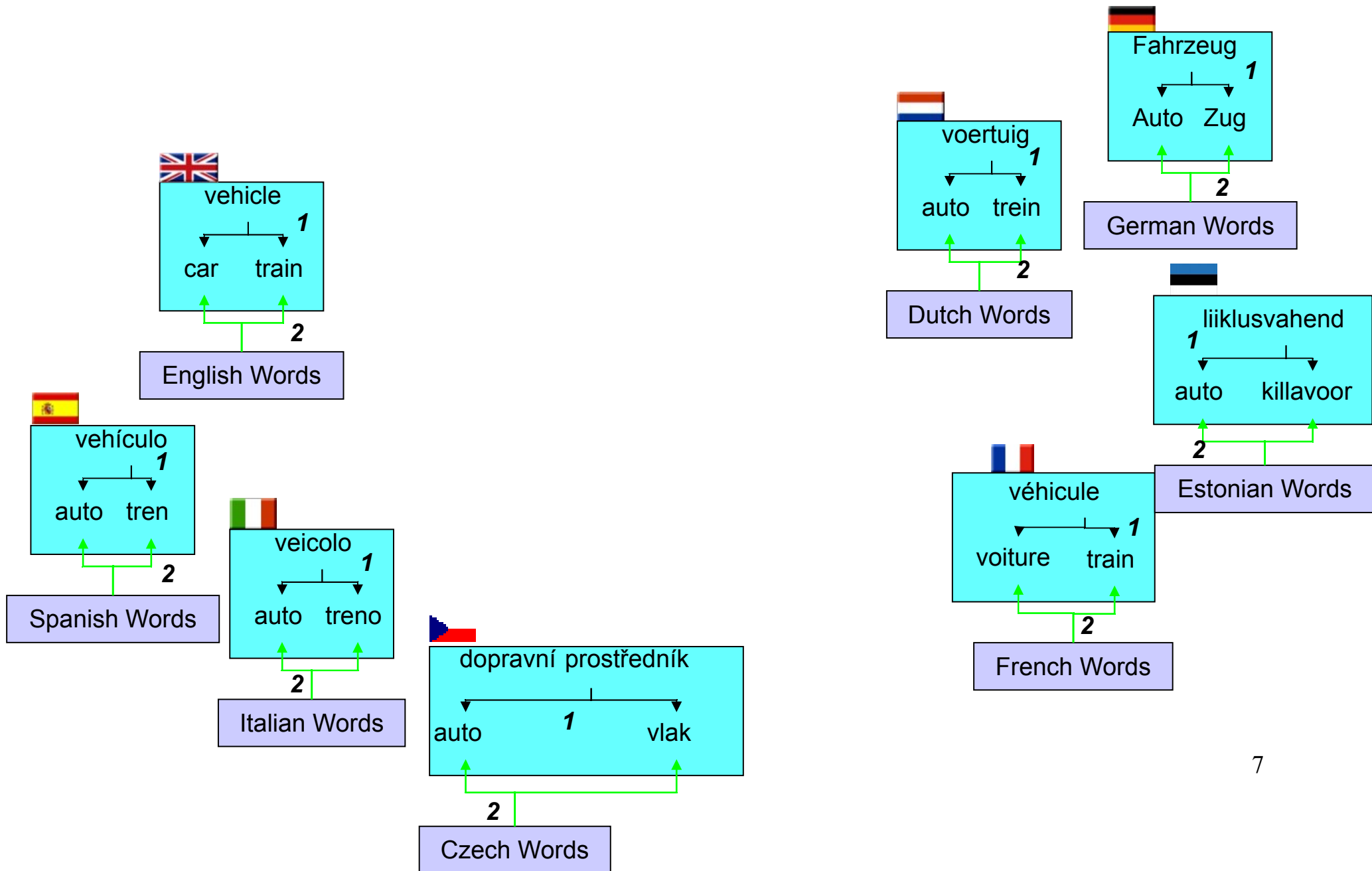
Wordnet Data Model



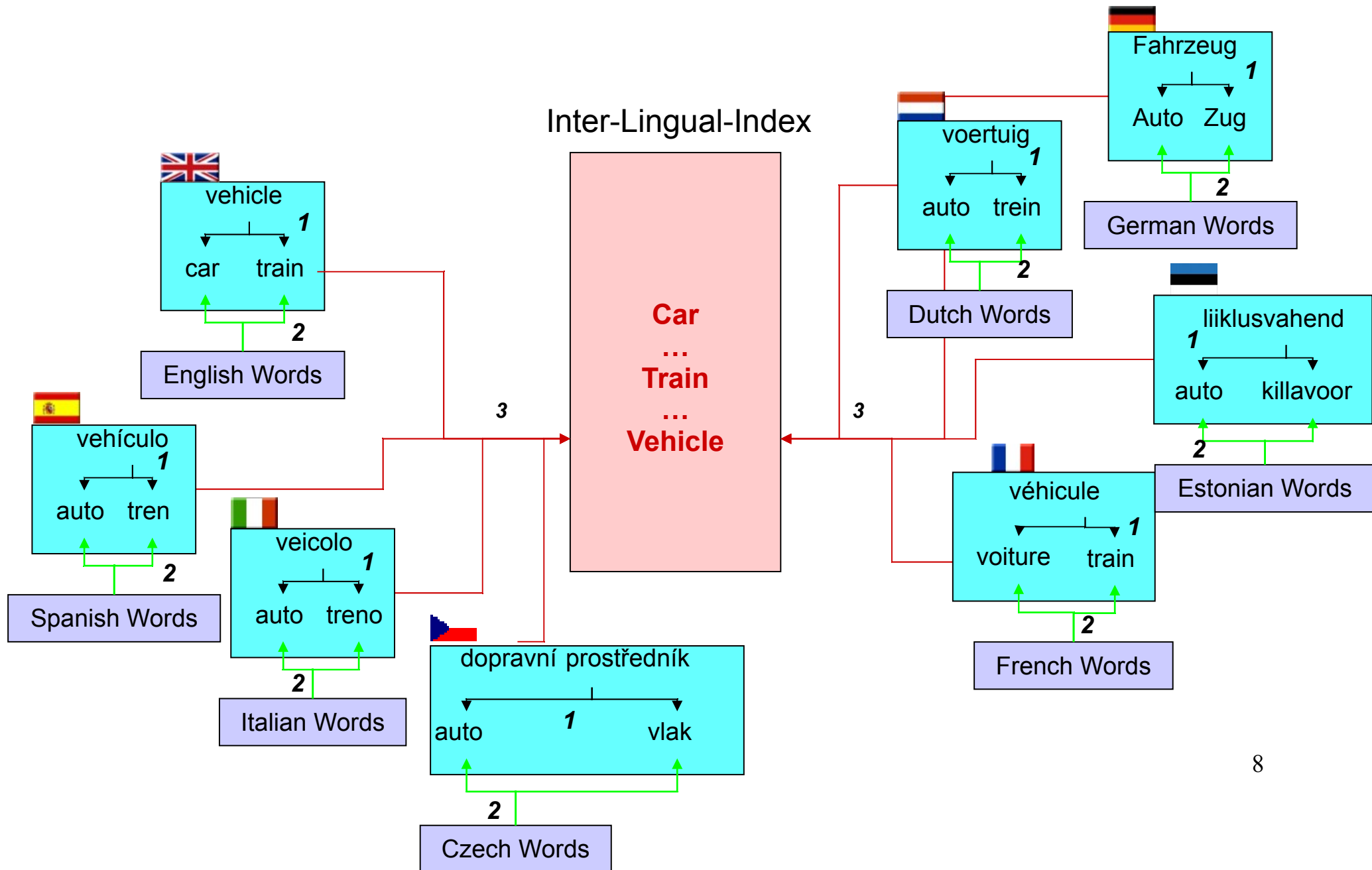
Cross-lingual wordnet model



Cross-lingual wordnet model



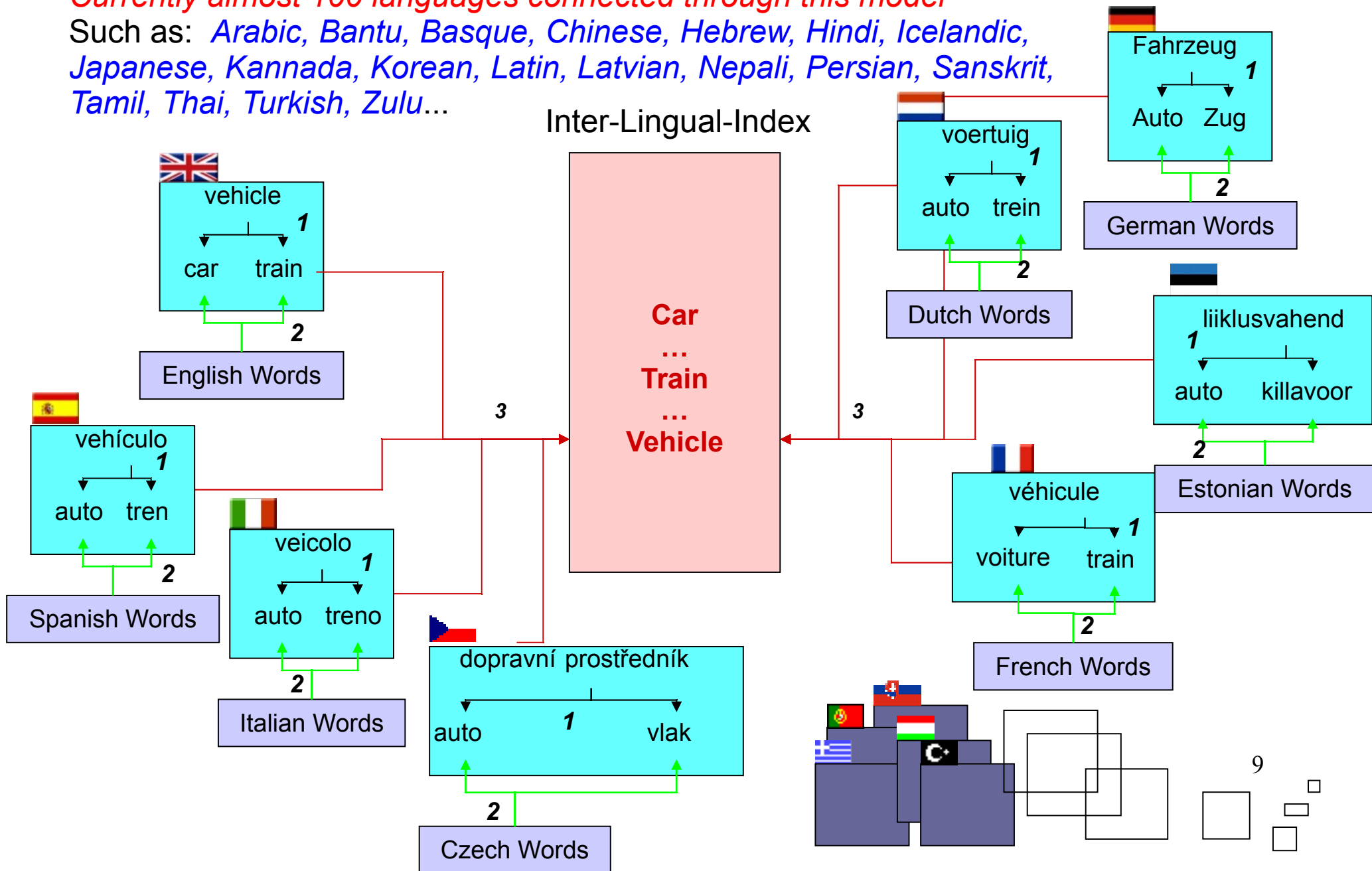
Cross-lingual wordnet model



Cross-lingual wordnet model

Currently almost 100 languages connected through this model

Such as: Arabic, Bantu, Basque, Chinese, Hebrew, Hindi, Icelandic, Japanese, Kannada, Korean, Latin, Latvian, Nepali, Persian, Sanskrit, Tamil, Thai, Turkish, Zulu...



Cross-lingual wordnet model

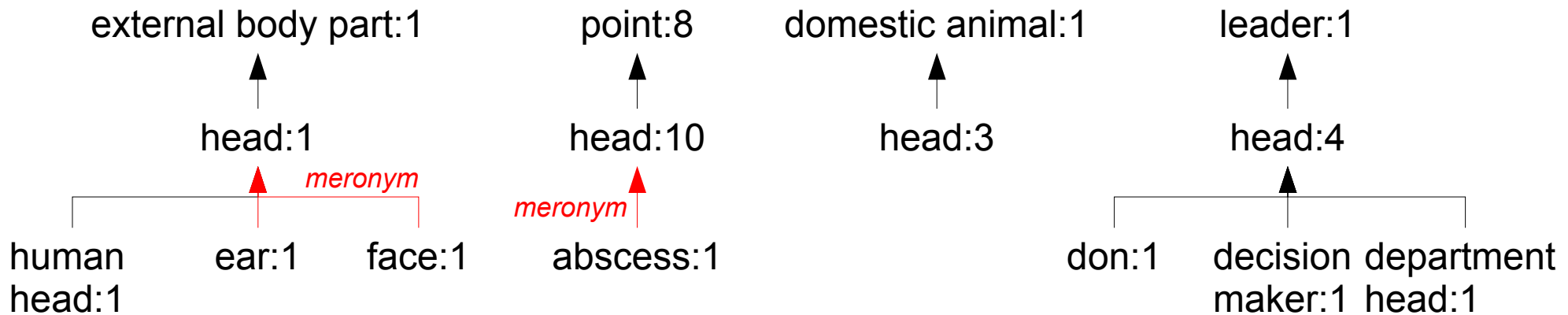
- Vocabularies of languages are unique networks of forms and concepts, both cultural and linguistic artifacts
- Vocabularies also show universal properties of their network structure
- Linking vocabularies to concepts provides unique opportunities for studying universals and idiosyncrasies of linguistic and conceptual structures

Network properties of wordnets

- Study by Sigman and Cecchi (2001):
 - Vertices are wordnet concepts for nouns (66K)
 - Edges are:
 - Semantic relations between concepts: hypernyms, meronyms
 - Form relations between concepts: polysemy, e.g. head (30), line (29), point (24)
- All relation sets show a scale-invariant distribution when correlating nr. of concepts with number of links (power-law behavior)
- Hypernym relations as a basis, while adding meronyms, polysemy and random links

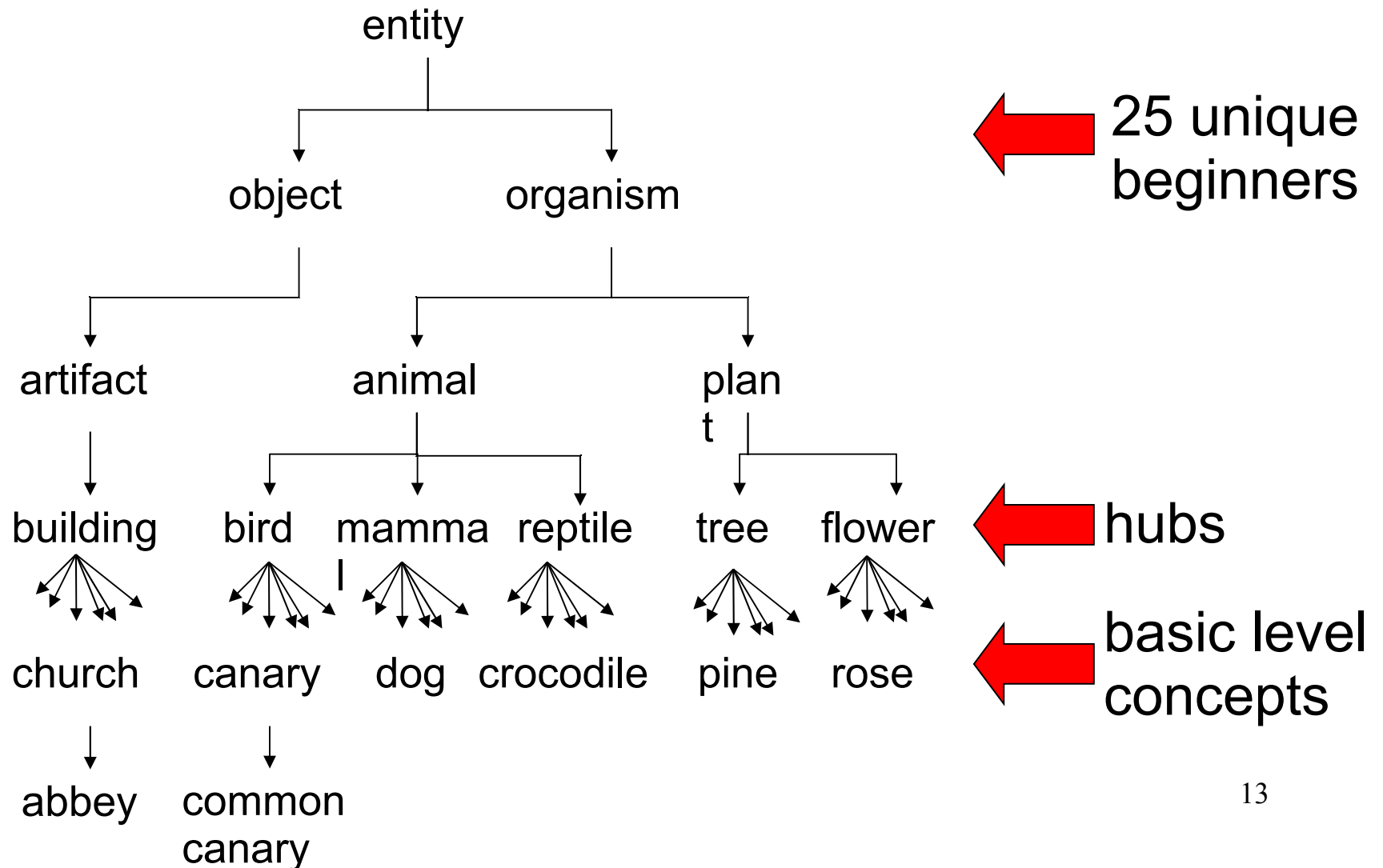
Network properties of wordnets

- Conclusion: adding polysemy creates a *smaller* world:
 - Reduces characteristic length (average min. distance) from 11.9 to 7.4 (semi-random is 8.5)
 - Increases the clustering from 0.0002 to 0.06
- What does it mean that most ambiguous words form small world hubs with underspecified meanings?



Overall structure of wordnets

Traditional cognitive model

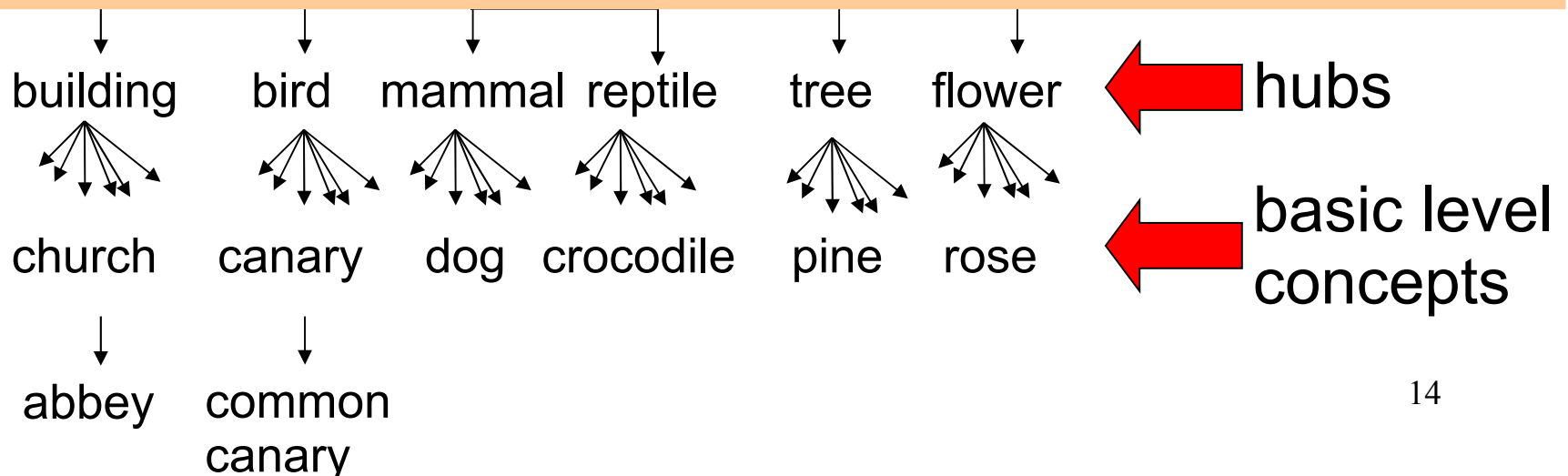


Overall structure of wordnets

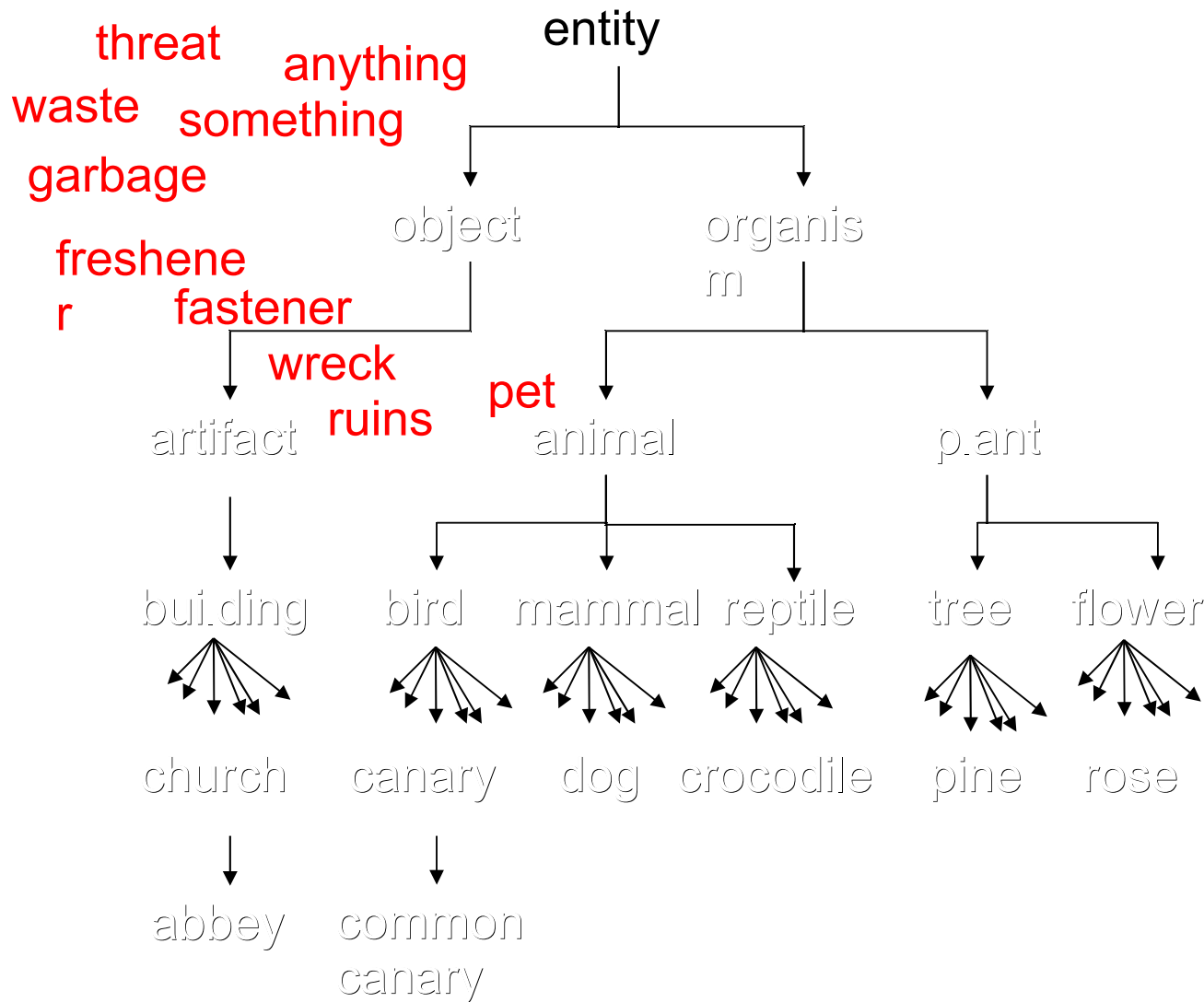
Traditional cognitive model

→ Basic level concepts

- Categories are formed as a balance of two principles:
 - predict most features
 - apply to most subclasses
- Level where most concepts are created
- Level that amalgamates most parts
- Most abstract level where you can still draw a picture



Overall structure of wordnets

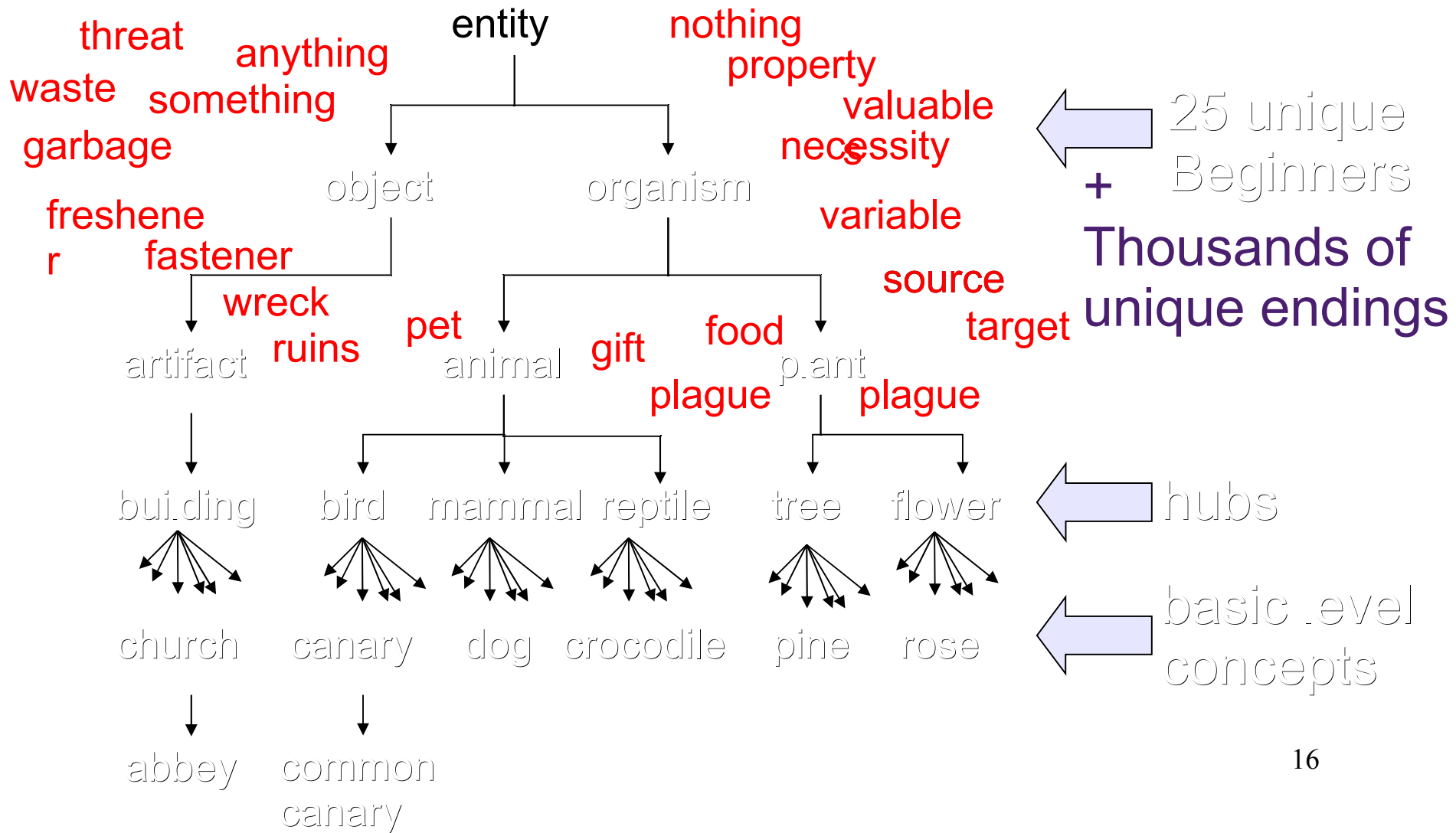


← 25 unique
+ Beginners
Thousands of
unique endings

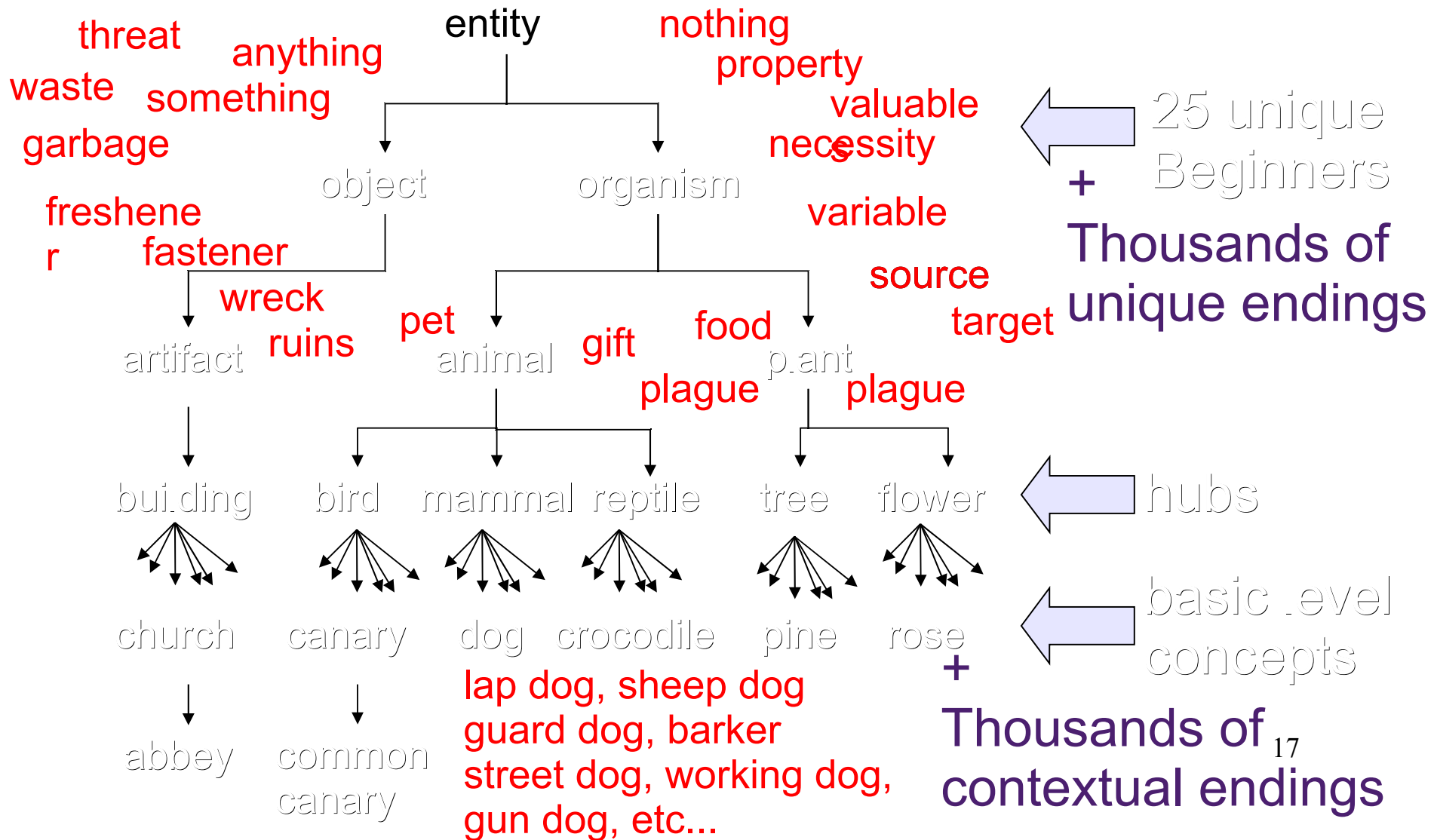
← hubs

← basic level
concepts

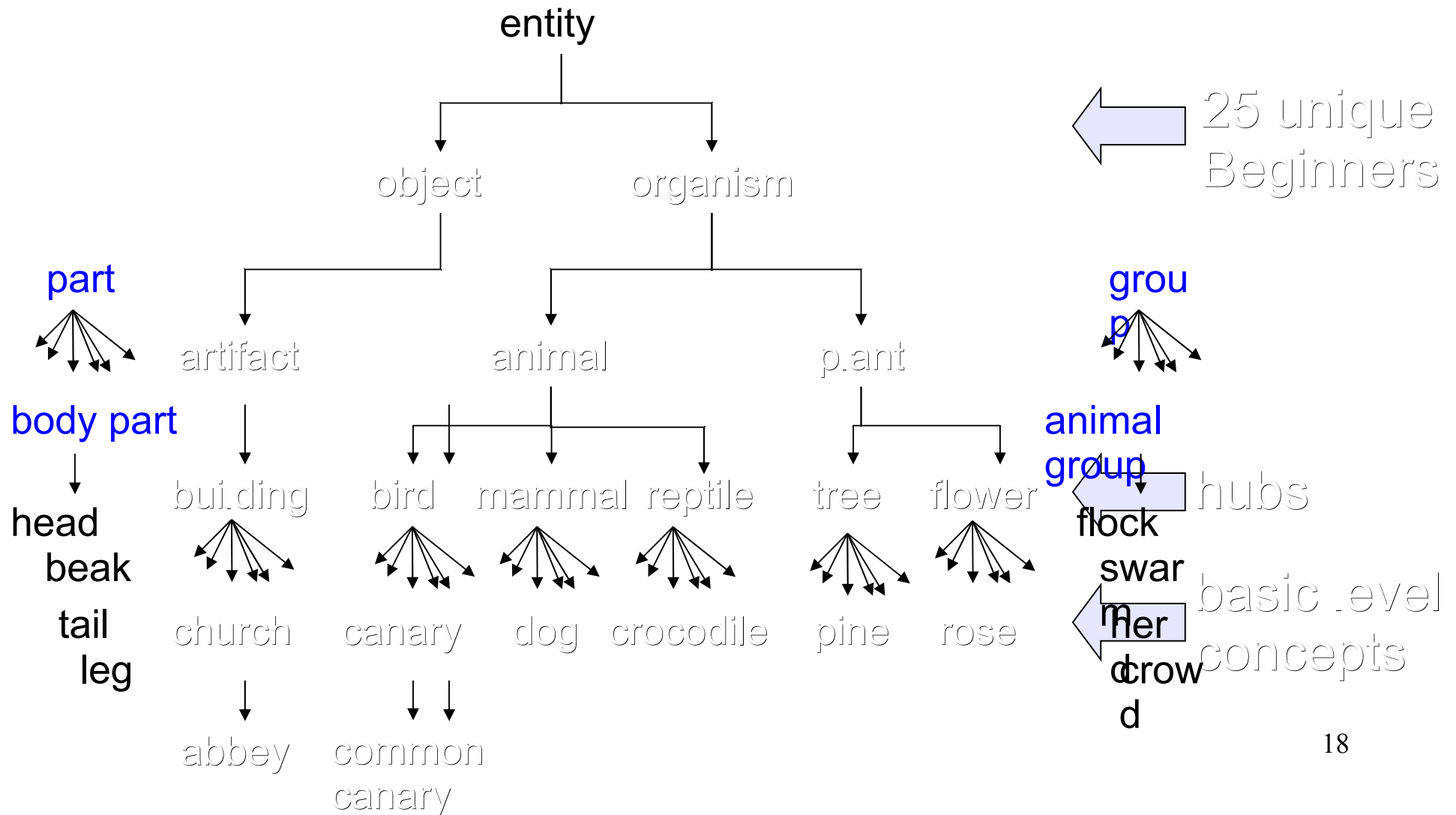
Overall structure of wordnets



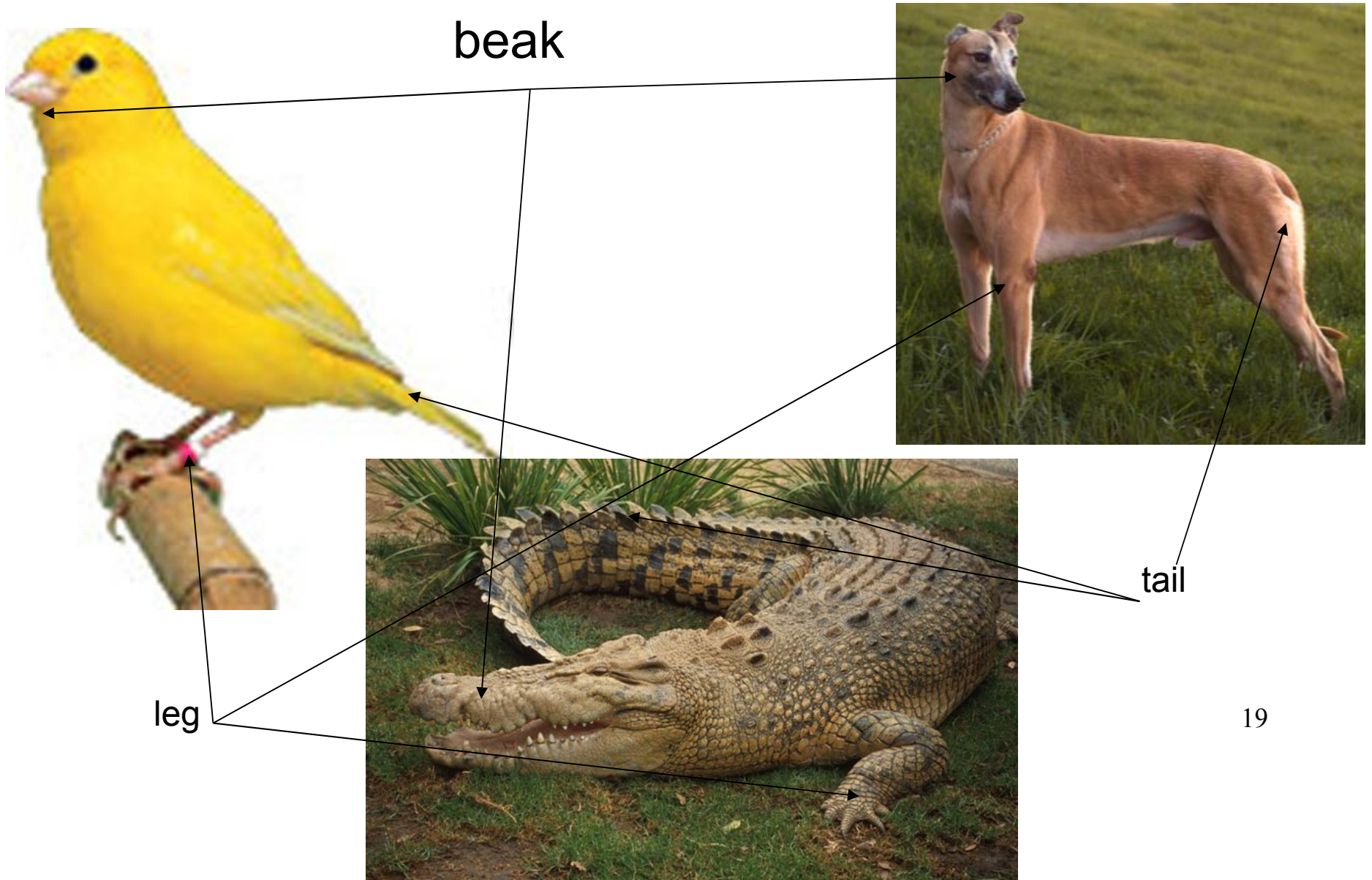
Overall structure of wordnets



Overall structure of wordnets



Ambiguity of words for parts



Ambiguity of words for parts



Why is language minimalistic?

- Language is both grounded in personal experience and social interaction
- Language mediates between knowledge of speaker and addressee and is never fully explicit
- Gricean maxim of quantity:
 - Make your contribution as informative as is required for the current purposes of the exchange, within a cultural setting.
 - Do not make your contribution more informative than is required for the purpose and the cultural background.
- We can do with less words to talk about many things

Conclusions

- Vocabularies of language form real-world networks in many ways
- The most frequent and ambiguous words represent hubs in small word structures, which are supposed to be most efficient and robust
- When we communicate, it is more efficient to use a smaller set of symbols whose meaning can be stretched
- Each language is an abstract symbolic structure, heavily influenced by the social and cultural carving and tuning in a language community

Network properties of vocabularies

- Most frequent words:
 - have most meanings,
 - are co-occurrence hubs,
 - are hubs in semantic networks incorporating ambiguity
- There are many more words than cognitive salient things: *lap dog, barker, street dog vs. poodle, Newfoundland*
- There are many words that can be applied to many different cognitive salient things: *friend, pet, threat, danger, menace, taxable object, purchase*, which can also refer to *dogs*
- Language is inherently vague and ambiguous as a real-world network but why?

Word networks of co-occurrence

- Study by Ferrer I Cancho & Solé:
- British National Corpus 460K/470K words and $1.61 \cdot 10^7$ $1.77 \cdot 10^7$ edges based on co-occurrence relations
- Network properties:
 - Average connections: 70/74 links
 - Cluster coefficient: 0.43/0.68, compare 0.00014 random
 - Average min. path length: 2.67/2.63, compare 3.06/3.03 random

Networks of word meanings

