Encoding Commonsense Lexical Knowledge into WordNet

Emanuele Pianta
joint work with Gianluca Lebani

Fondazione Bruno Kessler – HLT Group
University of Trento – Center for Mind/Brain Sciences

International Workshop on WordNet and Sketch Engine 9 March 2012

Overview

- Concept Feature Descriptions (FDs)
  - What are they, and how are they used in cognitive science and language rehabilitation
- Can we use the WN conceptual model to encode FDs?
- Is there a closed set of relations allowing WN to represent the common sense lexical knowledge contained in FDs?
- Can all content of FD be represented in WN?

Concept Feature Descriptions (FDs)

- Human generated concept-description pairs:

  `<dog> has 4 legs | barks | is not so big`

- Are considered by cognitive scientists as a window into human semantic memory (Cree et al., 2003)
- A long-lasting tradition of feature norms collection in cognitive psychology (since Rosch & Mervis, 1975; Mc Rae et al. 2005)
  - Subjects are presented with a set of concept names and asked to produce the features they think are important for each concepts
- Exploited in the treatment of anomic patients (cfr. Nickels, 2002)

In speech therapy practice, it is common to test noun-related knowledge in terms of associations between *target word* and *Feature Descriptions*

Category		**v/n**	**Target**	**Feature**	**Resp**
TOOL	funz	n	penna	si usa per girare la frittata	
TOOL	morf	v	chiodo	è appuntito	
FOOD	col	v	mela	è lucida e spesso rossa	
ANIMAL	morf	v	cane	ha grandi mammelle sul ventre	
ANIMAL	dim	v	scoiattolo	è un animale piccolo	
BIRD	dim	v	rondine	è un uccello piccolo	
ANIMAL	morf	v	cammello	ha piedi e mani muniti di dita	
FOOD	enc	n	caramella	è amara	
ANIMAL	col	v	rinoceronte	ha la pelle grigia	

Work by the speech therapist

- Typically, in preparing a task the therapist
  - exploits his/her semantic knowledge for finding stimuli
  - has to (manually) check on available resources
- E.g.: preparation of a semantic questionnaire
  - frequencies are checked in a frequency lexicon
  - <Concept> FD pairs are compiled by hand

    `<nail> has a pointed end`

    `<apple> is red`

    `<squirrel> is a small animal`
- Can a computational tool help the therapists?

STaRS.sys

- Semantic Task Rehabilitation Support System (FBK – UniTn)
  - helper for a therapist preparing a semantic task

- Retrieving concepts from specifications
  - E.g.: highly frequent animal concepts with highly distinctive colour features and a high mean feature distinctiveness
    - *output*: "zebra", "polar bear", "tiger", "leopard", …
  - Related task: *feature generation*
- Retrieving information associated to a concept
  - E.g.: perceptual features of concept "banana"
  - E.g.: functional features of concept "table"
  - Related task: *semantic questionnaire*
- Comparing concepts
  - E.g.: animals living in a similar/different habitat than "lion"
    - *output*: "leopard", "cheetah" ... vs. "seal", "gorilla" ...
  - Related task: *odd-one- out*

Can WordNet be used as a backbone for STaRS.sys?

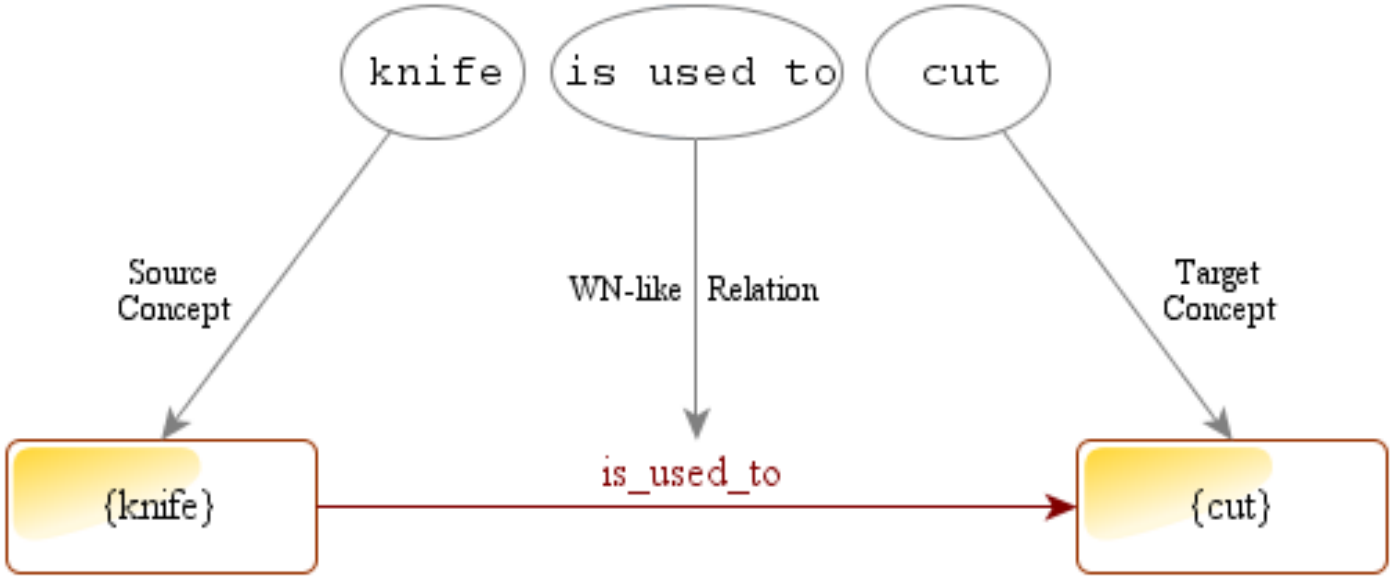
- Pros
  - Based on psycholinguistic assumptions
  - Easy to understand and use by therapists (wrt logics oriented formalisms)
  - Implements a full isa-hierarchy
- Issues
  - A. How can Feature Descriptions (FDs) be encoded in WN?
  - B. Does available WN relations cover what is needed to encode common-sense lexical knowledge?
  - C. Can all the content of Feature Descriptions be encoded in the WN conceptual model?

A. How to encode FDs into WN - 1

- A simple approach: append FDs to synset glosses (see usage examples)
  - PRO: easy to implement
  - PRO: can be useful for some WN applications
  - CONS: most usages of STaRS.sys (e.g. calculating concept similarity, retrieving concepts) require a more explicit representation of the semantic information contained in FDs

A. How to encode FDs into WN - 2

Representing a FD as a WN-like relation between a Source concept and a Target concept



BUT: what relations do we need to cover all Feature Descriptions?

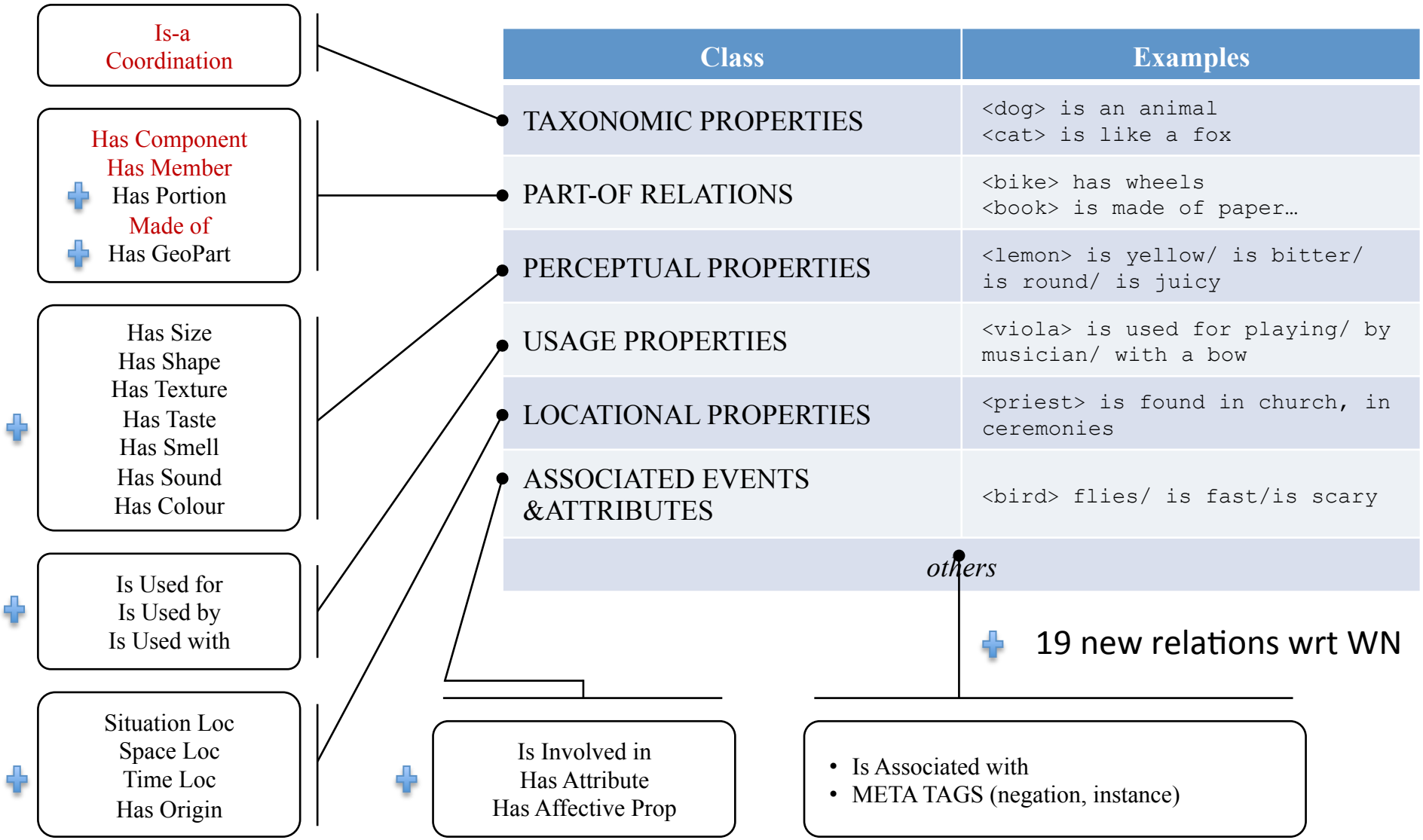
What WN relations are needed?

- A Feature Description classification is useful per se:
  - for selecting feature types of interest
  - for implementing most of the feature-based semantic measures available in the literature
- Further Advantage:
  - A complete FD classification will help us determine the range of semantic relations needed for encoding FDs
- Requirements of the classification:
  - Cognitive Plausibility
  - Intuitiveness
  - Robustness

Towards an exhaustive list of Feature Types / Relations

- Focus on first-order entities
  - i.e. concrete and physical entities ”(publicly) perceivable by the senses and located at any point in time, in a three-dimensional space” [Lyons (1977)]
- We started from relevant proposals :
  - used in a therapeutic context [e.g. CERIN, Laiacona et al (1993), Boyle & Coelho (1995)]
  - originated in cognitive psychology studies [e.g. Wu & Barsalou (2009), Cree & McRae (2003)]
  - motivated by well specified theoretical (ontological) explanations [e.g. Winston et al (1987)]
  - implemented in an extensive semantic resource [e.g. Fellbaum (1998), Lenci et al (2000)]
- NB: As a matter fact, it turns out that we can establish a *one-to-one* mapping between Feature-types and WN relations

A proposal for the classification of Feature Types / Relations



Class	Examples
TAXONOMIC PROPERTIES	<dog> is an animal <cat> is like a fox
PART-OF RELATIONS	<bike> has wheels <book> is made of paper…
PERCEPTUAL PROPERTIES	<lemon> is yellow/ is bitter/ is round/ is juicy
USAGE PROPERTIES	<viola> is used for playing/ by musician/ with a bow
LOCATIONAL PROPERTIES	<priest> is found in church, in ceremonies
ASSOCIATED EVENTS &ATTRIBUTES	<bird> flies/ is fast/is scary
others	

Evaluation of the classification

- A first version of the classification has been evaluated through inter-subject agreement (Lebani & Pianta, 2010)
- 5 non-expert Italian speakers (University students)
- asked to annotate 300 concept-feature pairs from a non-normalized version of the collection by Kremer et al (2008)
- Inter-subject agreement: Fleiss' Multi-π = 0.73
- We take the resulting agreement as a measure of both reliability (i.e. reproducibility) and usability/learnability
- A slightly modified version of the classification has been evaluated with therapist, producing comparable (or better) results.

C. Can all a FD content be encoded as a triple {synset} relation {synset}?

- To answer this question we run an experiment in two steps:
  - 1. A new collection of FDs related to 50 concepts
  - 2. Systematical encoding of all the FDs of 5 concepts

Step 1. A new collection of FDs - *why*

- Why not using existing Feature norms?
  - lack of coverage for certain types of feature
  - due to the organization of the semantic memory
  - due also to the methodology exploited for eliciting descriptions?
  - due also to the normalization procedure?
- as opposed to our need to cover the largest and most varied set of semantic aspects as possible

Step1. A new collection of FDs - *how*

- participants: 60 Italian speakers (students - researchers)
- same concepts as Kremer & Baroni (2011):
  - 50 concepts from 10 categories: bird, body part, building, clothing, fruit, furniture, implement, mammal, vegetable, vehicle
- task: to describe 10 concepts by answering to a list of questions based on the semantics of our relations
  - E.g.: What is the *color of* Cherries; what *kind of* cherries are there, etc.
- every concept has been described by 12 subjects with the help of an on-line questionnaire

FDs collected (vs. Kremer norms)

- Raw Descriptions: 18,884 (vs. 8,250 in Kremer norms)
  - 377.68 descriptions per concept (vs. 170.4)
  - every subject produced in the average 31.47 descriptions per concept (vs. 4.96)
- Preprocessing:
  - 1,023 (5.4%) descriptions were deleted (technical, wrong or autobiographical infos)
  - 2,247 (11.9%) description were assigned to different types

Step 2: Encoding new FDs into MWN

- criteria:
  - Some amount of interpretation cannot be avoided, but reduce as much as possible the need for it
  - Whenever possible, do not simplify / reduce the content of FDs
- outcome: StarsMultiWordNet a dedicated version of the Italian MultiWordNet
- preliminary results with 5 concepts
  - seagull, finger, chair, corn, airplane
  - 1,785 raw descriptions

Issues in the encoding phase

- Feature Description Normalization
- Ambiguity
- Loose talk
- Complex concepts
- Negation
- Cardinality
- Certainty

Feature Description Normalization

- All Feature Description collections undergo a normalization phase in which equivalent FDs are merged:
  - String-wise identical
  - Syntactic variant
  - Semantically equivalent (synonym expressions)
- However in most cases the notion of semantic equivalence is not well def.
- In our case you used WordNet synsets as synonymy criterion
  - `<wheel> is a component of a car`
  - `<wheel> is an auto part`
  - equivalent because they can be both mapped into a meronymic relation linking {wheel} and {car, auto}
- 1,785 raw equal or variant descriptions reduced to 871 relation instances
- 59 semantically equivalent descriptions have been merged into 29 relation instances

Ambiguity

- Lemmas contained in FDs (representing target concepts) can be ambiguous (can be assigned to more than one synset)
- In our sample, an av. of 3.2 synsets per lemma
- A procedure has been designed that allows the encoder to decide whether to create a relation with only one of the synsets or all of them
  - <cherry> grows in {gardens}/{grounds}
  - <corn> can be found in a {basement, cellar}/{root_cellar, cellar}
- 64 descriptions (7.3% of the sample) have been encoded with more than one relation

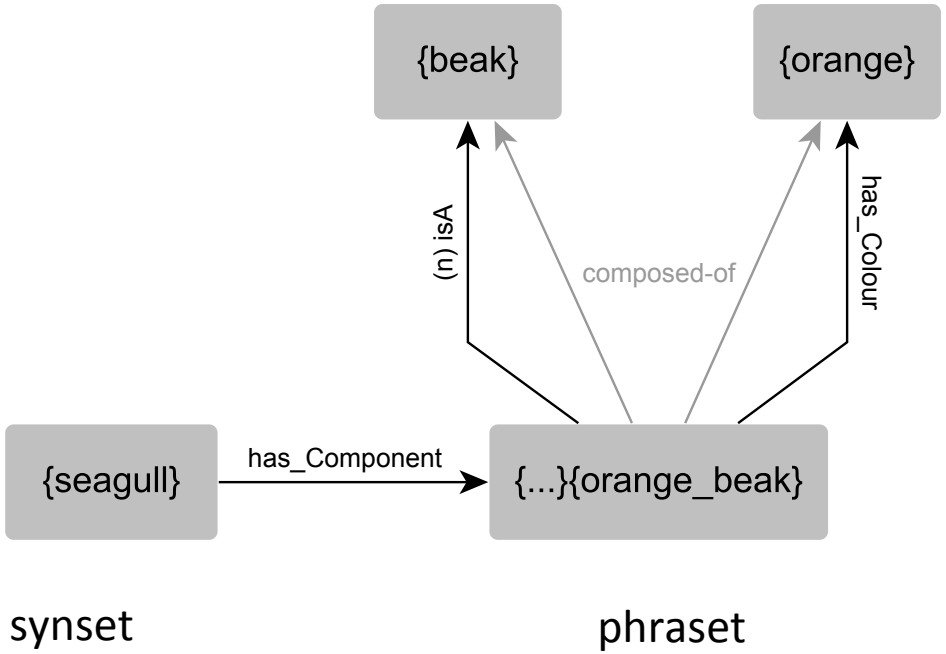
Loose talk

- Subjects may ignore some terms or may simply not remember them when they produce the FD. As a consequence, certain free descriptions contained in FDs could be rephrased by using a specific lexical unit:
- E.g.: is used by people who cook.
- WN glosses can be used as a basis for mapping the free phrase into a synset
  - {cook} defined in the gloss as "*someone who cooks food*"

Complex concepts

- In some other cases however a concept is expressed in the FD by a free description that has no lexical correspondent,
- e.g. <seagull> has an orange beak

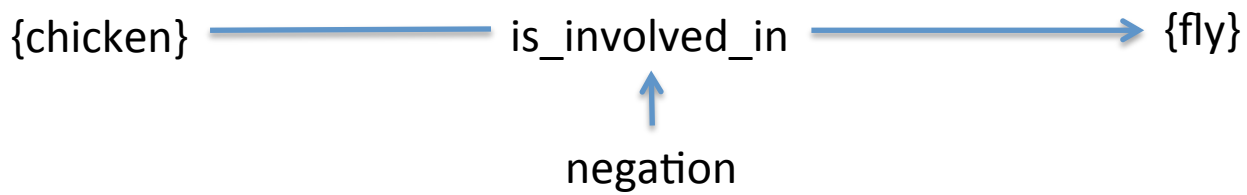
Phraset: a set of synonym free expressions
(as opposed to lexical units).
Can be used to represent the content of a lexical gap, or an alternative way of expressing a lexical unit
(Bentivogli e Pianta, 2004)



Negation

- McRae handles negative FDs as a specific Feature type (<bike> doesn't have an engine and <chicken> cannot fly go in the same class!)

- We follow the EWN way, by introducing in MWN relation features (also labels)



- As expected, features negated by subjects can be seen as blocking "expected" undesired implications.

Cardinality

- Many solutions have been proposed, but none of them is useful for our purposes.
- As an example, in Vinson and Vigliocco (2008), descriptions such as `has 4 wheels` are split into the two indipendent concepts `4` and `wheels`.
- Again, relation labels are the solution

has_cardinality:4,6

NB: cardinality can be expressed as a range

Certainty

- In standard FD normalization modifiers such as “generally”, “most of the times”, “sometimes” are ignored.
- We propose a new relation label, called Certainty, representing the intuition of the language speaker about how strong is his/her expectation that a certain relation holds between the instances of two concepts

Certainty cont.

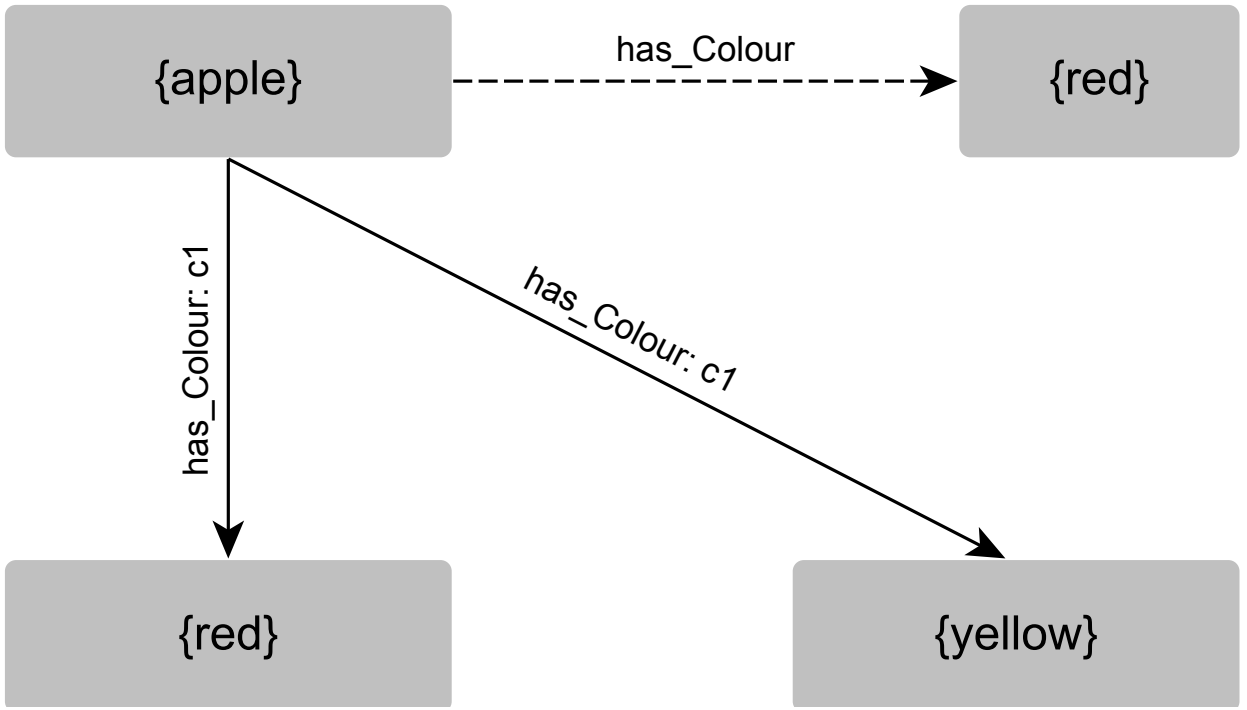
- *True by definition*: the speaker thinks that the relation between two concept instances holds because of how the concepts are conventionally defined; no exceptions are admitted:
  - E.g.: <cat> is a feline.
- *Certain*: the speaker expects the relation to hold unless an anomaly occurs, which needs a causal explanation:
  - E.g: <man> has arms.
- *Probable*: the speaker expects the relation to hold most of the times; however if this does not occur it is not perceived as an anomaly.
  - E.g. <wardrobe> is typically made of wood.
- *Possible*: the speaker expects the relation to occur sometimes, but not most of the times.
  - <wardrobe> can be made of plastic.

Conjunction and disjunction

- Given two relation instances of the same type what is the logical relation between them?
- We need to define a default for each relation type
- E.g. by default
  - *part_of* relations are in conjunction
    (a tree has roots and branches)
  - *has _color* relations are in disjunction
    (apples are either red or yellow)
- However, in a few cases we need to overcome the default
  - some apples are red and yellow

Conjunction and disjunction cont.

Specific logical relations (over-writing default relations) are represented through relation labels (à la EWN)



Some results

- The semantics of 795 normalized FDs (91.3% of the total) could indeed be fully encoded as a semantic relation between two simple synsets.
- In 137 cases (15.7%) a synset for the focal concept of the description was missing.
- The encoding of 71 FDs required the creation of one or more phrasets.
- In 32 cases a part of the information expressed by the FD has been discarded.
- Only 5 raw descriptions were discarded because an efficient way to encode them was not found
  - e.g. “partially black”, “is high as half a person”

Conclusions

- An extended version of WordNet including a larger set of relations (+19), and a richer data structure (phrasets, relation labels) can be used to represent the vast majority of the information contained in Feature Descriptions.
- Only a small percentage of FDs cannot be represented through the extended WN conceptual model.

Thank you!

Concepts in Cognitive Sciences

- an abstraction?
- a definition?
- a logic formula combining semantic primitives?
- a set of postulates (logical implications)?
- a prototype?
- a mental image?
- a bunch or relations with other concepts?
- a list of features?

Concepts in Computer Science

- *KR Frame*: an isa relation + slots and facets (e.g. KL-ONE)
- *Synset*: A set or synonyms + relations with other concepts (WordNet)
- *FrameNet Frame*: an event and its typical participants
- *Ontological Concept*: a formally defined structure allowing for logical inference
- *Vector Space Models*: a set of word co-occurrences