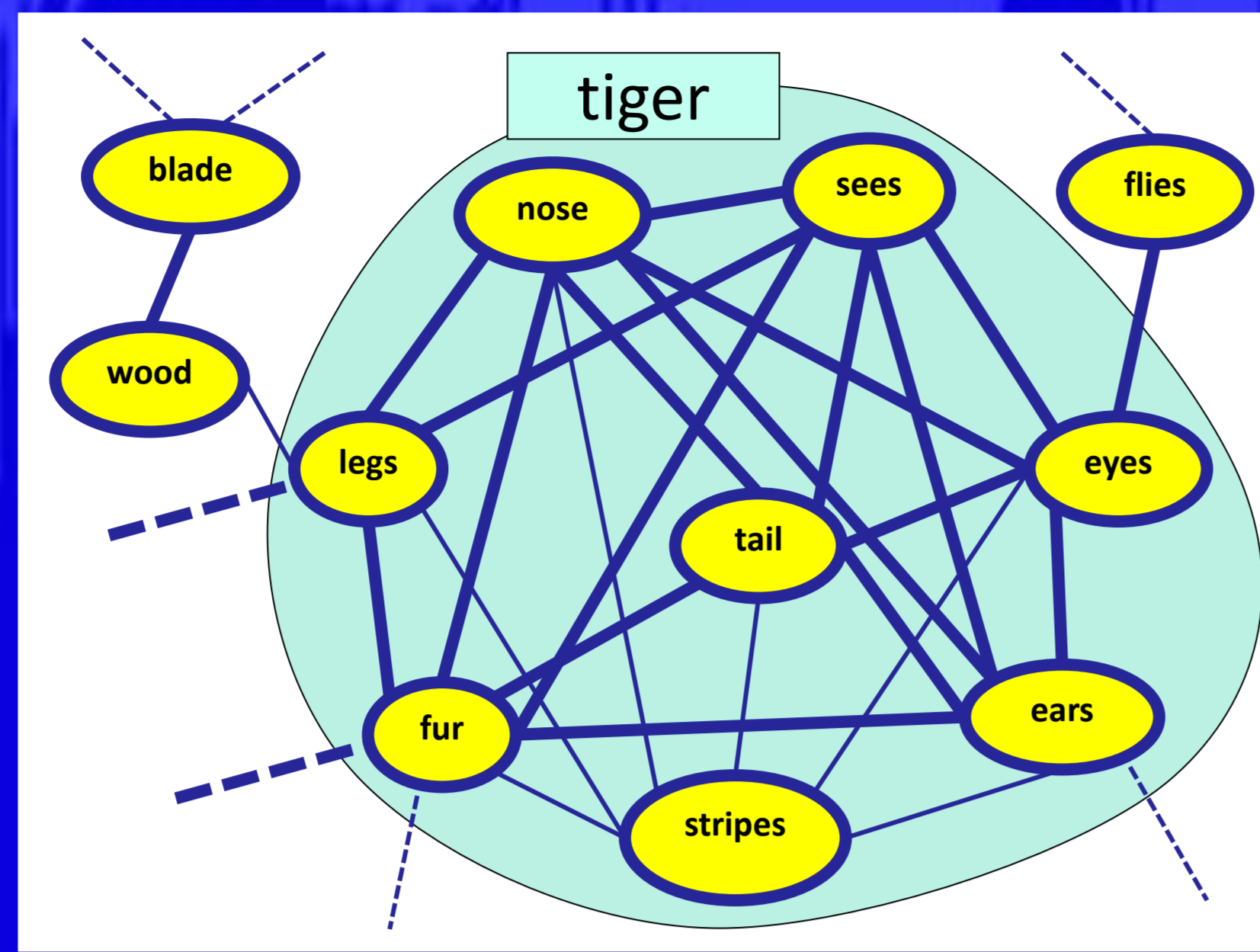


# The acquisition of unrestricted feature-based conceptual representations from corpora

## Background & Motivation

- Many theories of conceptual knowledge assume a distributed, feature-based representational framework
- Concepts are proposed to be patterns of activation across these feature units
- Question:** How can we estimate the knowledge people represent in this distributed, feature-based system?
- Traditional approach** – property norm data (e.g. McRae et al., 2005)
- Participants enumerate properties of concepts; researchers normalise the responses into frequency-weighted lists of features
- Problems with property norms**
  - Expensive and time-consuming to gather
  - Less salient (shared) features under-represented (e.g. *tiger* - *has eyes*)



concept	feature norm	freq
tiger	has_stripes	22
tiger	a_carnivore	17
tiger	an_animal	15
tiger	has_teeth	15
tiger	lives_in_jungles	14
tiger	has_eyes	N/A

## Computational Feature Extraction

- Idea** – Extract features automatically from corpora (e.g. Almuhaireb & Poesio, 2005; Barbu, 2008; Baroni et al., 2009, 2010)
- Very challenging task (unconstrained)
- These approaches have typically used lexico-syntactic pattern matching (c.f. Hearst, 1992)
- These approaches:
  - typically do not make use of deeper syntactic analysis (i.e. dependency parses)
  - do not focus on extracting relations linking the concept term and feature term
    - has, lives\_in, used\_for, etc.*

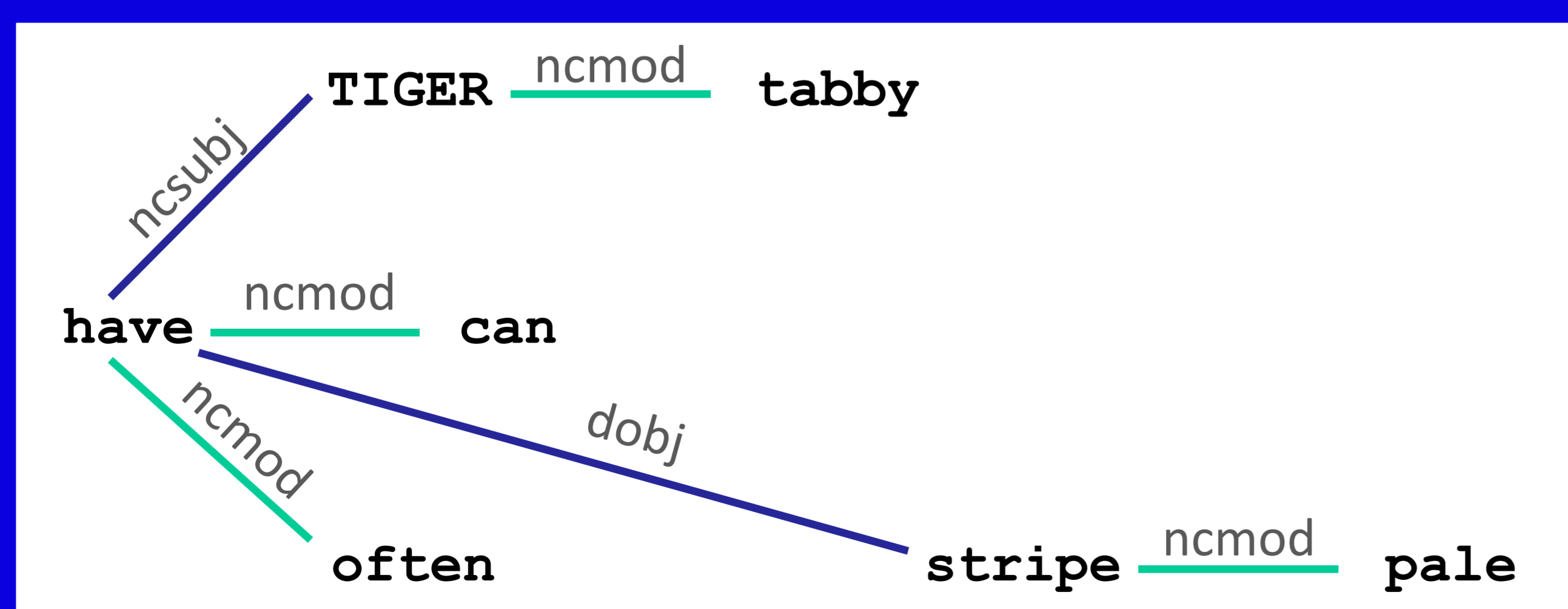
**Pattern:** *the <F> of a <C>*  
**Matched text:** "...the layers of an onion are..."  
**Extracted feature tuple:** <onion, layers>

## Our approach

- Use dependency parsing to extract property norm-like features from corpora
  - <concept, relation, feature> triples
- No constraints on <relation> (any verb) or <feature> (any adjective or noun)
  - e.g. <flute produce sound>, <deer have antler>, <swan be white>
- Use 3 kinds of information to guide extraction: encyclopedic, syntactic, semantic
  - encyclopedic: use encyclopedia corpus (Wikipedia)
  - syntactic: use dependency parsing (RASP) to guide extraction of candidate features
  - semantic: use analysis of norms to filter out poor-quality candidate features

## Candidate feature extraction

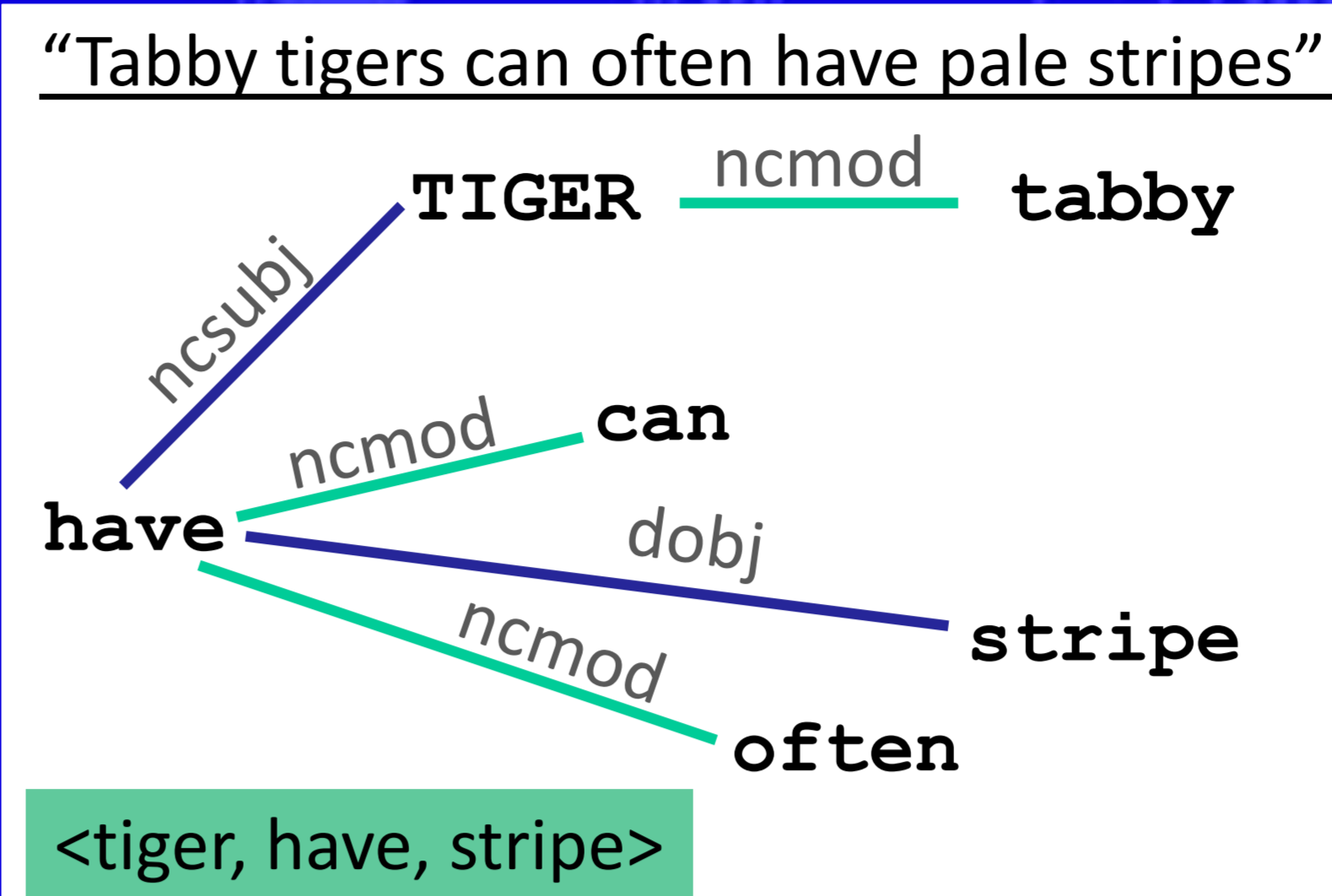
- Corpus:** Wikipedia (651M words)
- Parsed with RASP (Briscoe, 2006) and grammatical relations (GRs) were extracted
- Our aim: extract features for 500 concepts in the McRae et al. (2005) norms
- Stage 1. Candidate feature extraction**
  - Construct a graph with words as nodes and GRs as edges
  - Identify paths through the graph
  - If there is a verb in the path, extract the potential feature triple <concept, verb, feature>
- Example: "Tabby tigers can often have pale stripes"



- Extracted candidate triple: <tiger, have, stripe>

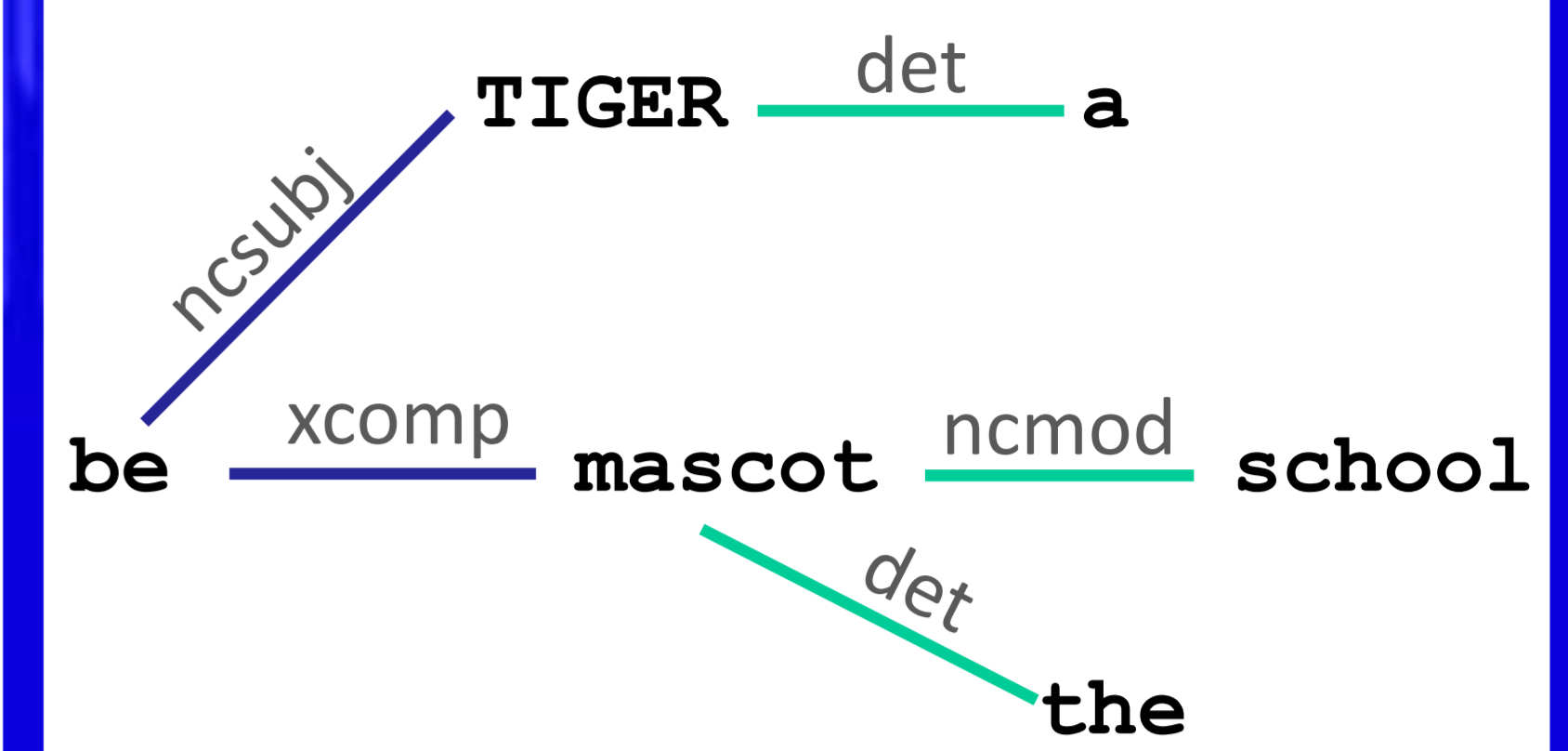
## Re-ranking & Filtering

- Candidate feature extraction uses syntactic information only, and can extract both correct and incorrect triples:



<tiger, have, stripe>

"The school mascot is a tiger"



<tiger, be, mascot>

- Stage 2. Semantic filtering**
  - Filters out poor quality feature triples
  - Based on conditional probabilities of different types of features and concepts occurring together
  - Calculated for clusters of McRae features and concepts obtained using Lin (1998) similarity metric
  - We re-weight extraction frequencies by the conditional probabilities

- Example conditional probabilities for clusters

		Concept clusters		
		Reptiles	Fruit/Veg	Vehicles
Feature clusters	Body parts	0.164	0.031	0.023
	Plant parts	0.009	0.130	0.014
	Events/Activities	0.100	0.060	0.140

- Example of reranking:

triple	rank before	rank after
<tiger have stripe>	91 <sup>st</sup>	23 <sup>rd</sup>
<tiger be mascot>	1 <sup>st</sup>	2632 <sup>nd</sup>

- Impact:** removes implausible features and increases the weight of correct shared features not typically found in property norms (e.g. <tiger have eyes>)

## Evaluation

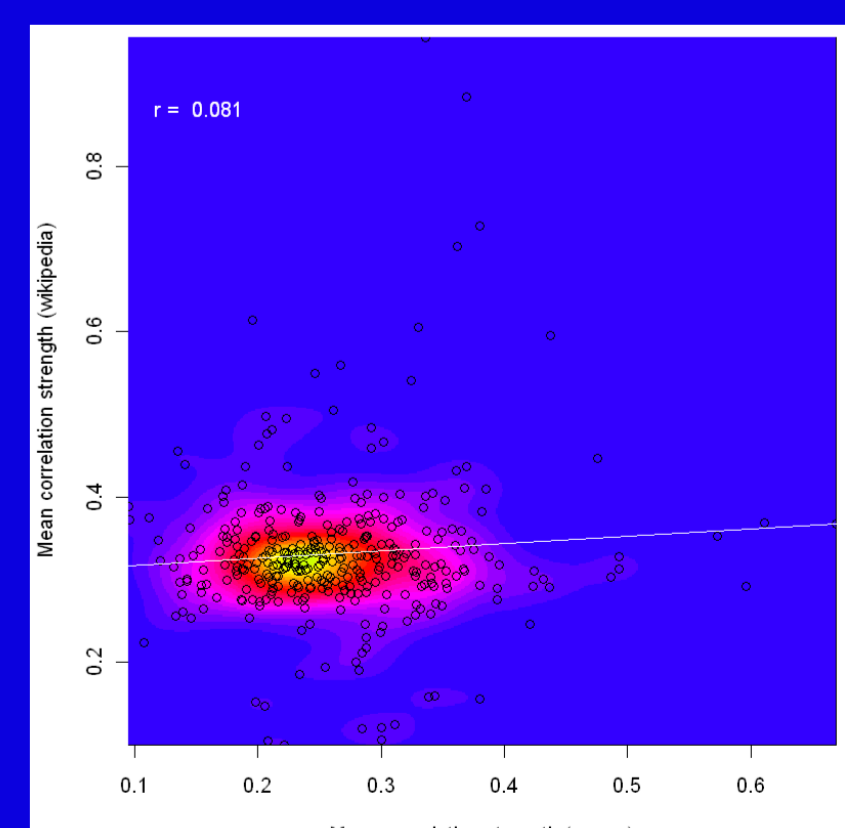
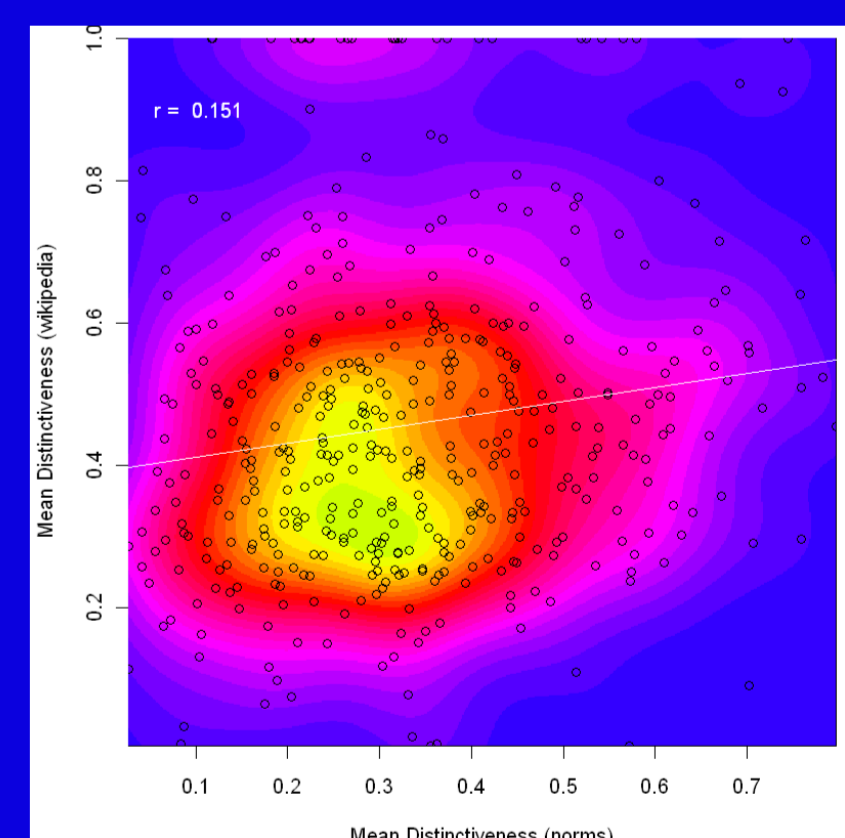
- Examples: top 12 reranked features for three McRae concepts:

cranberry		skunk		clarinet	
include	product	be	mammal	be	instrument
associate	dinner	include	predator	play	instrument
be	sauce	include	mammal	be	key
be	wine	be	animal	be	reed
associate	sauce	include	animal	use	instrument
be	food	keep	pet	play	musician
garnish	meat	be	white	score	symphony
soak	meat	be	pet	play	band
mix	juice	be	predator	be	music
drink	juice	be	black	include	instrument
use	food	be	fur	be	horn
be	sweet	spray	dog	be	musical

- Evaluation against the McRae Norms** (44 concepts in ESSLI set (Baroni et al. 2008))
  - Precision:** proportion of extracted features in the norms
  - Recall:** proportion of norms extracted by the model
- Problem: features can be true even when not in the norms
  - recall is more important than precision

Method	Prec.	Recall
SVD (baseline)	0.014	0.280
Extraction method - unfiltered	0.007	0.808
Extraction method - top 25%	0.018	0.626
Extraction method (reranked) - top 25%	0.023	0.685

- Evaluation in terms of conceptual structure:**
  - A method of evaluation that avoids direct comparison to property norms
- Conceptual structure variables (CSVs) describe the structure of concepts (Tyler & Moss, 2001)
  - Previously calculated on property norms (e.g. McRae et 2005, Randall et al 2004)
  - e.g. Number of Features, Mean Distinctiveness of features, Mean Correlation Strength of features (for each concept)
- We find significant (but weak) correlations with McRae norms for several CSVs



## Conclusions

- Syntactic, semantic and world knowledge can be used to guide the extraction of norm-like features from corpora
- Scope for further improvement:
  - constrain base extraction further (identify the most useful paths in the GR graphs)
  - enhance semantic filtering (derive semantic constraints from other sources)
  - Larger/broader corpora (e.g. combine Wikipedia and BNC)
- Improve evaluation (e.g. use behavioural and brain imaging data)