Taxonomies in a corpus: let's go for a ride!

Stijn Storms

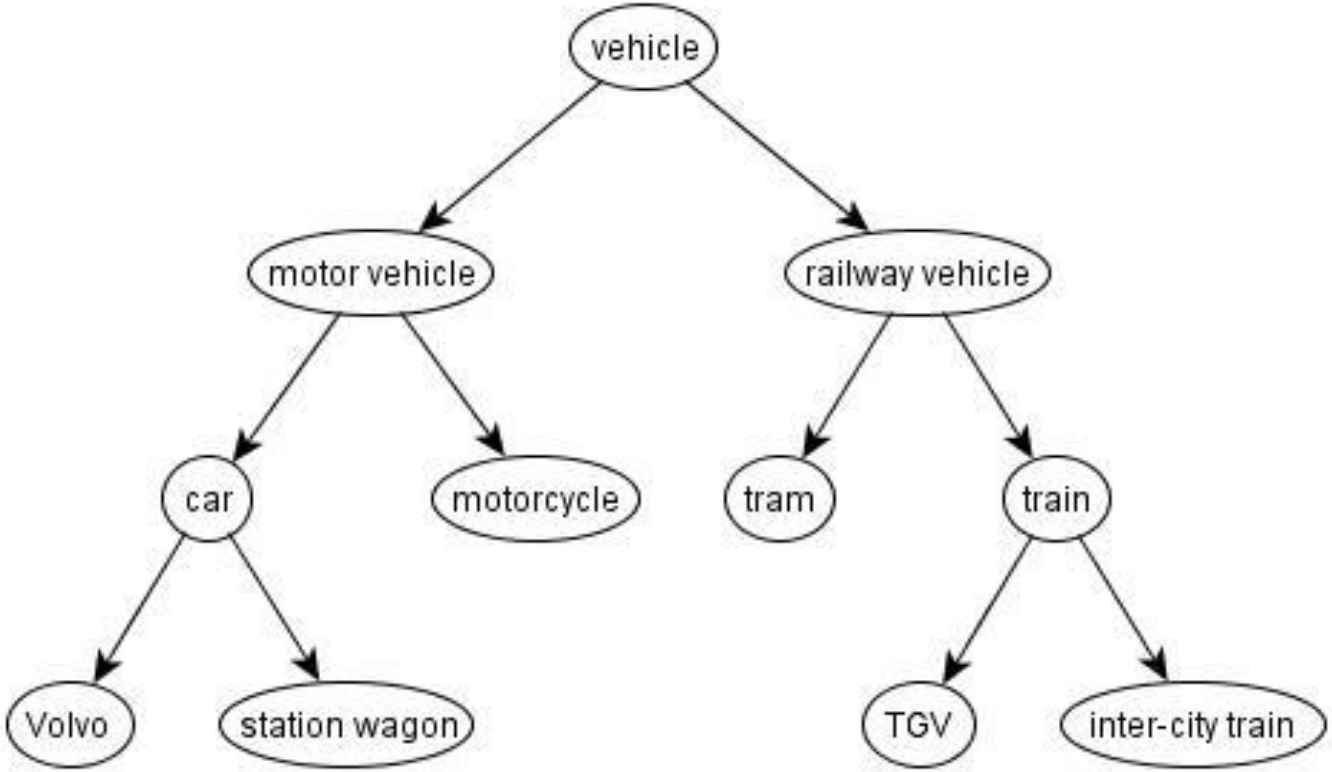
Dirk Speelman & Dirk Geeraerts

University of Leuven
RU Quantitative Lexicology and Variational Linguistics

Overview

- Research question
- From informativeness to corpus linguistics
- Methodology
- Case study

Taxonomy


vehicle
motor vehicle
railway vehicle
car
motorcycle
tram
train
Volvo
station wagon
TGV
inter-city train

Basic level

= a cognitively preferenced level by which we think about any one thing

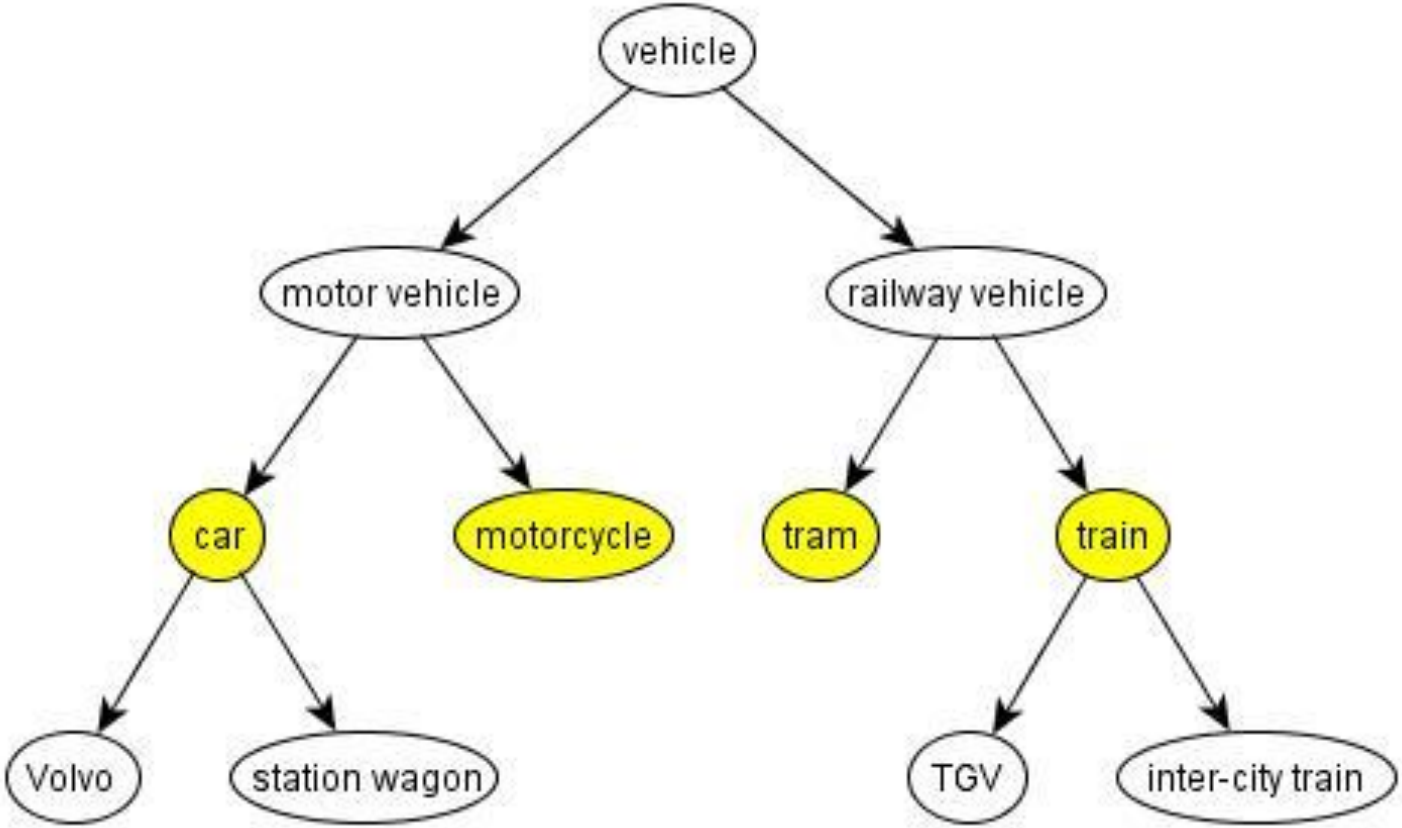
- Linguistics

  Berlin, B., Breedlov.De and P. H. Raven (1973). "GENERAL PRINCIPLES OF CLASSIFICATION AND NOMENCLATURE IN FOLK BIOLOGY." American Anthropologist **75**(1).

- Psychology

  Rosch, E., C. B. Mervis, W. D. Gray, D. M. Johnson and P. Boyesbraem (1976). "BASIC OBJECTS IN NATURAL CATEGORIES." Cognitive Psychology **8**(3): 382-439.

Basic level


vehicle
motor vehicle
railway vehicle
car
motorcycle
tram
train
Volvo
station wagon
TGV
inter-city train

Research setting

We aim at studying the unique nature of basic level concepts from a corpus linguistic point of view. We are specifically interested in uncovering distributional patterns, as captured by, among others, vector space models.

Research setting

Geeraerts, D., S. Grondelaers and P. Bakema (1994). The structure of lexical variation : meaning, naming, and context. Berlin ; New York, M. de Gruyter.

Research question

When we look at instances of basic level concepts and their subordinates, do we find that, on average, the subordinate ones show more mutual contextual similarity than the basic level ones?

Research question

THE DAILY NEWS

www.dailynews.com THE WORLD'S FAVOURITE NEWSPAPER - Since 1879

Alarming 50% of population takes car for 200 metres

A bicycle is a mechanism that has transported man for many decades. Nowadays, bicycling is a great way to move from one place to another in short distances. I believe you should use a bicycle for short distances to protect the environment as well as your health. The purpose of this essay is to explain why people should ride a bicycle for short distance trips.

By riding a
decrease ga
improve the
lives, and pr
environmen
The first rea
for short dis
can cause a
pollution. B
people use t
short distan
the rates of
more polluti
environmen

THE DAILY NEWS

www.dailynews.com THE WORLD'S FAVOURITE NEWSPAPER - Since 1879

Drugs found in car

Two children were in a car in which police found more than $1 million worth of drugs on Sunday.
Police allegedly found more than a million dollars worth of drugs in a car during a traffic stop in Albany on Sunday night.
Police spokeswoman Susan Usher said traffic officers pulled over a vehicle near the Albany Airport just before 10am.
"A search of the vehicle located methylamphetamine with an estimated street value worth more than $1 million," she said.

The Department of Child Protection was notified that children were present and took them into its care.
Mark Anthony Stolban, 30, and Tahnee Rochelle Hill faced Albany Magistrates Court this morning but were not required to enter pleas.
They face charges of possession of amphetamine and possession of cannabis with intent to sell or supply and are expected to face court again on Thursday.

Research question

THE DAILY NEWS

www.dailynews.com THE WORLD'S FAVOURITE NEWSPAPER - Since 1879

BMW caught for speeding

A head gamekeeper who drove at 100mph has escaped a ban after claiming he was being "pushed" along the road by an Audi driver.

Calum Sharp, 41, claimed he was pressurised into putting his foot down in his BMW X5 because another motorist was travelling at high speed behind him.

Perth Sheriff Court was told that officers charged him with clocking 114mph even though it was the vehicle behind him which was caught on the speed gun.

Fiscal dep
told the co
carrying ou
when they
overtaking
M90 south
Solicitor K
defending
slightly un
the speed
vehicle the
by the han

THE DAILY NEWS

www.dailynews.com THE WORLD'S FAVOURITE NEWSPAPER - Since 1879

Ford and BMW involved in violent collision

PALM HARBOR — One person was killed in a traffic crash involving at least two cars early Wednesday on McMullen-Booth Road, the Florida Highway Patrol said.

The crash was reported about 5:30 a.m. near Curlew Road.

Authorities said a 2008 BMW sedan, driven by Jonathan Peter Bytautas, 29, of Clearwater, ran a red light at the intersection and slammed into an eastbound Chevy pickup. Both cars then collided with a westbound Ford truck, driven by George Neil Paajanen, 60, of Palm Harbor.

During the crash, one of the cars slammed into a light post and another caught on fire.

The driver of the Chevy, Robert Henry Raymer Jr., 54, of Holiday, was pronounced dead at the scene.

Bytautas sustained serious injuries; Paajanen sustained minor injuries.

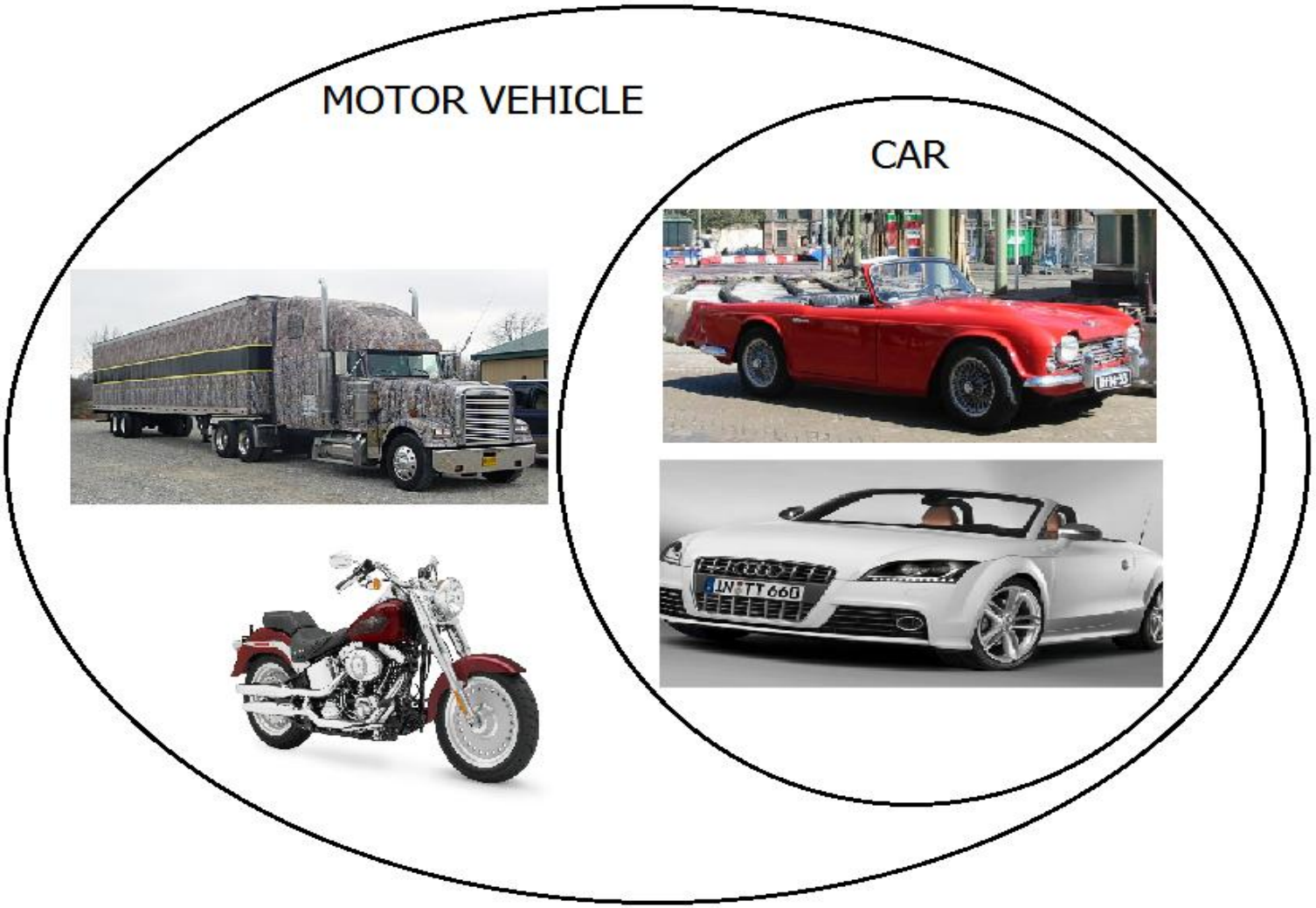
The westbound lanes of Curlew Road and the northbound lanes of McMullen-Booth Road were blocked as troopers were on scene.

An investigation remained under way Wednesday morning.

Overview

- Research question
- From informativeness to corpus linguistics
- Methodology
- Case study

Informativeness



Informativeness

Category	Possible hierarchical feature set
vehicle	mobile machine, transport passengers or cargo
car	mobile machine, transport passengers or cargo, has a mobile engine, has seating, has four wheels
sports car	mobile machine, transport passengers or cargo, has a mobile engine, has seating, has four wheels, designed for spirited performance

Informativeness

Informativeness refers to the amount of information which is associated with concepts.

Informativeness is thought to go up when we go down in the taxonomical tree.

Informativeness

Can this psychological notion of informativeness give rise to the discovery of patterns in the distribution of terms as we observe them in a corpus ?

Terminological translation

Informativeness

When we go down in the taxonomy, concept members tend to be more similar

term occurrences

Individual term occurrences

*Advances such as the driverless **car** are no longer the stuff of sci-fi*

*It seems the more modern the **car**, the more difficult it is to service it's needs*

*I ate McDonald's in my **car***

Terminological translation

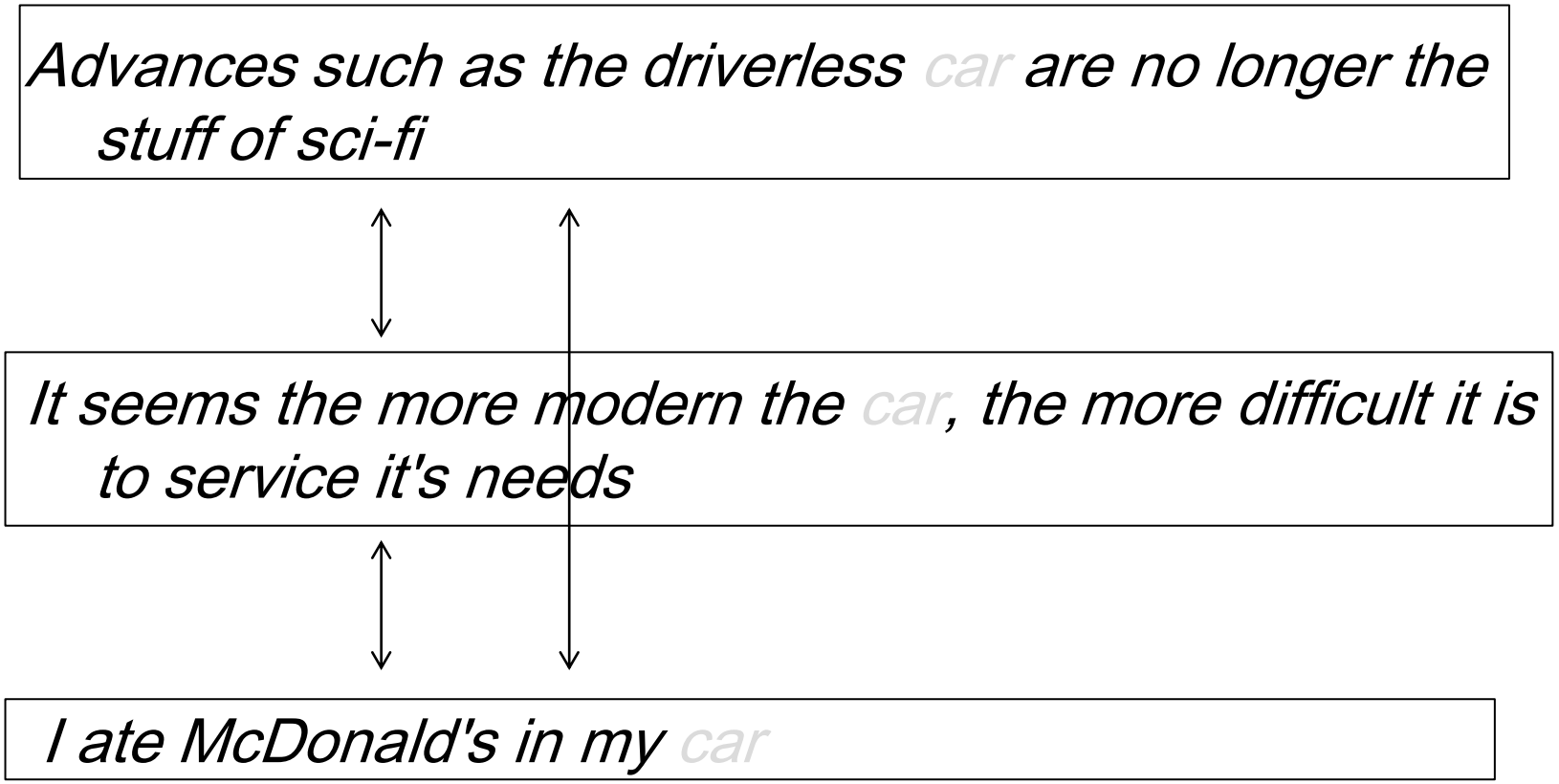
Informativeness

When we go down in the taxonomy, concept members tend to be more similar

You shall know a word by the company it keeps (Firth, 1957)

context similarity

Individual term occurrences



Hypothesis

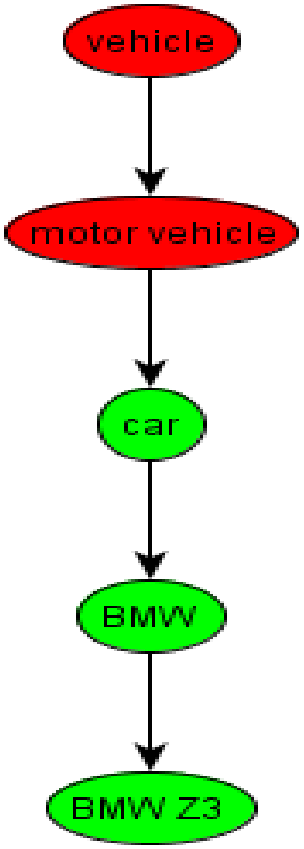
When going down in the taxonomy, term occurrences tend to show up in increasingly similar contexts

Hypothesis

When going down in the taxonomy, term occurrences tend to show up in increasingly similar contexts

But is this what we would expect from a linguistic point of view? What do we know about the linguistic usage of terms from different taxonomical levels?

Hypothesis



Hypothesis

Basic level terms constitute an inherent neutral level of specificity

*My **car** is parked right outside*

<-> *My **BMW** is parked right outside*

<-> *My **means of transportation** is parked right outside*

Cruse, D. A. (1977). "The pragmatics of lexical specificity." Journal of linguistics **13**: 11.

Hypothesis

Subordinate terms are useful when there is a domain that contains many members of a basic-level category that need to be distinguished

*The accident involved a green **BMW**, a grey **Volvo S40** and a silver **Opel Zafira**.*

Murphy, G. L. (2002). The big book of concepts. Cambridge, Mass., MIT Press.

Hypothesis

Put in a simplified way, basic level terms consitute a ‘default’ choice, which we can expect in a wide range of contexts. Their lower-ranked alternatives however are reserved for ‘special’ circumstances, so that we can expect to see them in a more restricted set of contexts

Hypothesis

Put in a simplified way, basic level terms consitute a ‘default’ choice, which we can expect in a wide range of contexts. Their lower-ranked alternatives however are reserved for ‘special’ circumstances, so that we expect to see them in a more restricted set of contexts

within-term similarity$_{\text{BASIC LEVEL TERM}}$

< within-term similarity$_{\text{SUBORDINATE TERM}}$

Hypothesis

No studies found on usage differences between subordinate and sub-subordinate terms

within-term similarity$_{\text{SUBORDINATE LEVEL TERM}}$
< within-term similarity$_{\text{SUB-SUBORDINATE TERM}}$?

Overview

- Research question
- From informativeness to corpus linguistics
- Methodology
- Case study

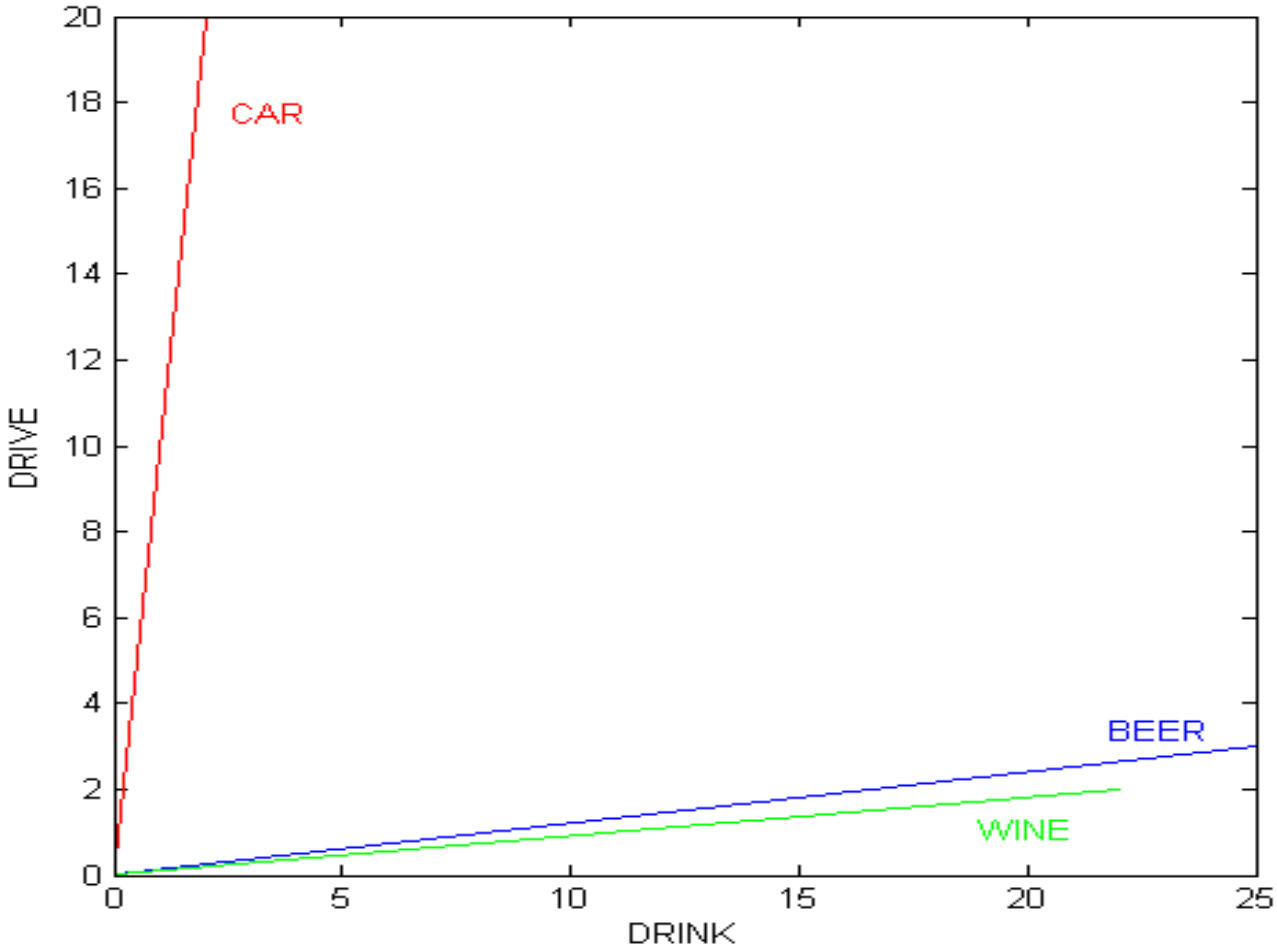
Vector space modelling

Vector space modelling is based on the idea that similarity of context hints at semantic similarity. It allows us to define and measure a distributional form of similarity between linguistic targets of our choice, e.g. between terms, between documents, …

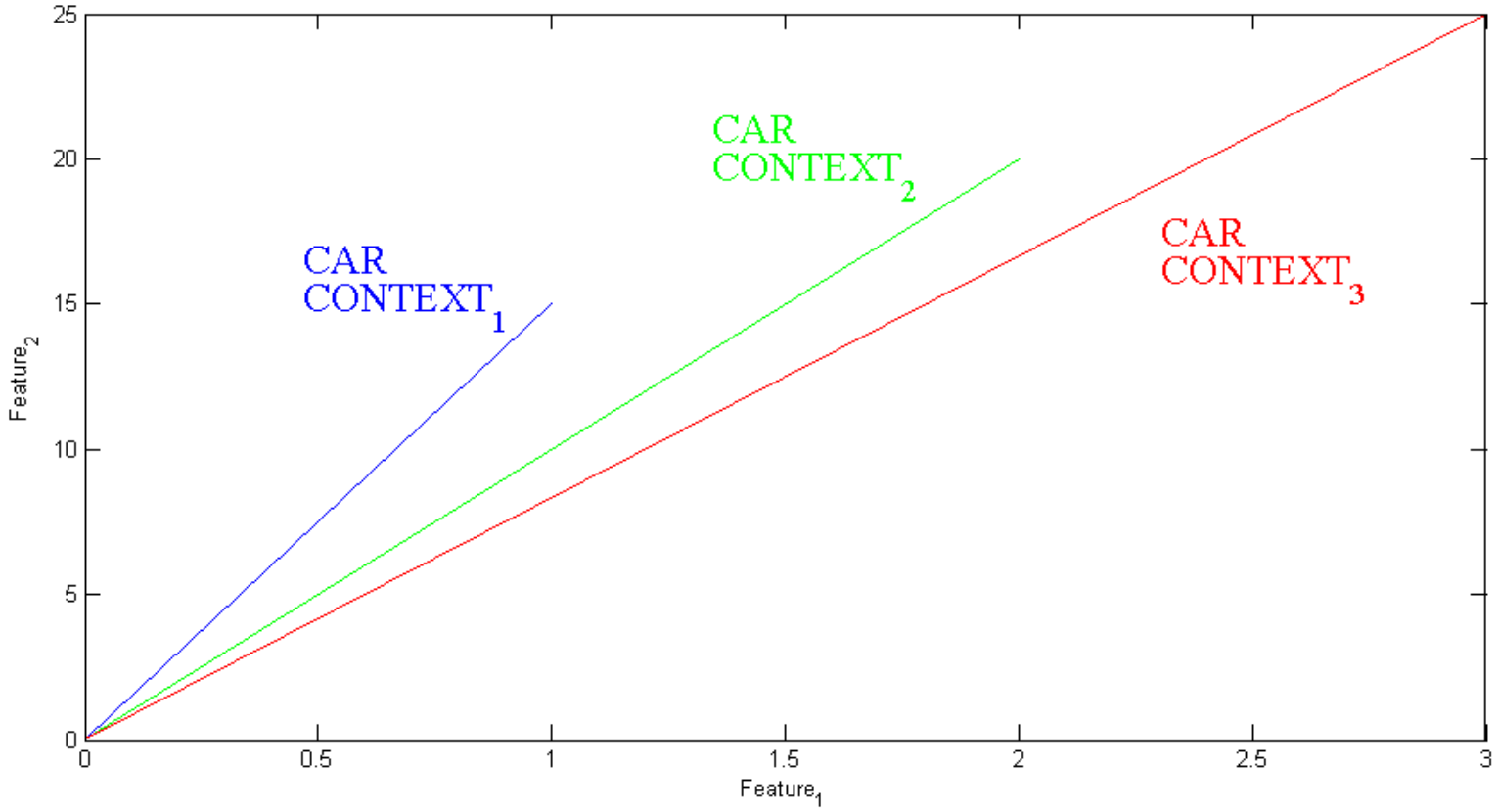
Vector space modelling

	DRINK	DRIVE	...
car	2	20	
beer	25	3	
wine	22	2	

Vector space modelling



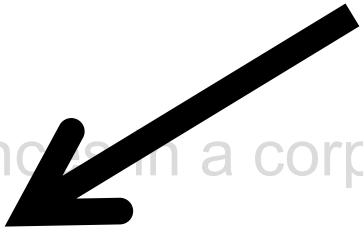
Individual term occurrences



Recipe for within-term similarity

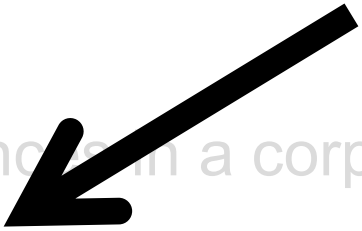
1. Gather its occurrences in a corpus
2. For each occurrence
    1. Select its neighbouring context words
    2. Replace each context word by a precomposed co-occurrence vector
    3. Add these co-occurrence vectors together to get its context vector
3. Calculate similarities between these context vectors and take the average

Recipe for within-term similarity



1. Gather its occurrences in a corpus
2. For each occurrence
   1. Select its neighbouring context words
   2. Replace each context word by a precomposed co-occurrence vector
   3. Add these co-occurrence vectors together to get its context vector
3. Calculate similarities between these context vectors and take the average

Recipe for within-term similarity



1. Gather its occurrences in a corpus
2. For each occurrence
   1. Select its neighbouring context words
   2. Replace each context word by a precomposed co-occurrence vector
   3. Add these co-occurrence vectors together to get its context vector
3. Calculate similarities between these context vectors and take the average

Context word selection

I ate McDonald's in my car

Context word selection

I ate McDonald's in my car

Context word selection

I ate McDonald's in my car

Context word selection

I ate McDonald's in my car

Context word selection

I ate McDonald's in my car

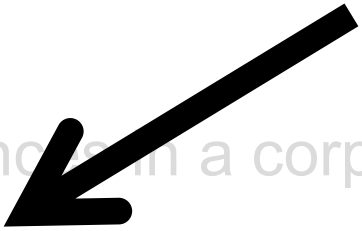
Context word selection

I ate McDonald's in my car

Context word selection

I ate McDonald's in my car

Recipe for within-term similarity



1. Gather its occurrences in a corpus
2. For each occurrence
   1. Select its neighbouring context words
   2. Replace each context word by a precomposed co-occurrence vector
   3. Add these co-occurrence vectors together to get its context vector
3. Calculate similarities between these context vectors and take the average

Dealing with data sparseness

I ate McDonald's in my car

I had breakfast in my car

Dealing with data sparseness

I ate McDonald's in my car

I had breakfast in my car

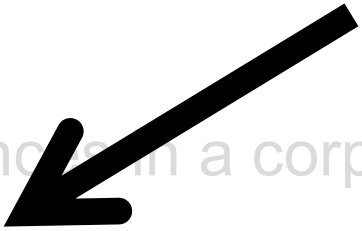
	EAT	MCDONALD'S	BREAKFAST
context1	1	1	0
context2	0	0	1

Dealing with data sparseness

Use precompiled matrix which contains information about the distribution of words as a whole

	...	YOGURT	SANDWICH	...
...				
McDonald's	...	13	20	...
breakfast	...	59	102	...
...				

Recipe for within-term similarity



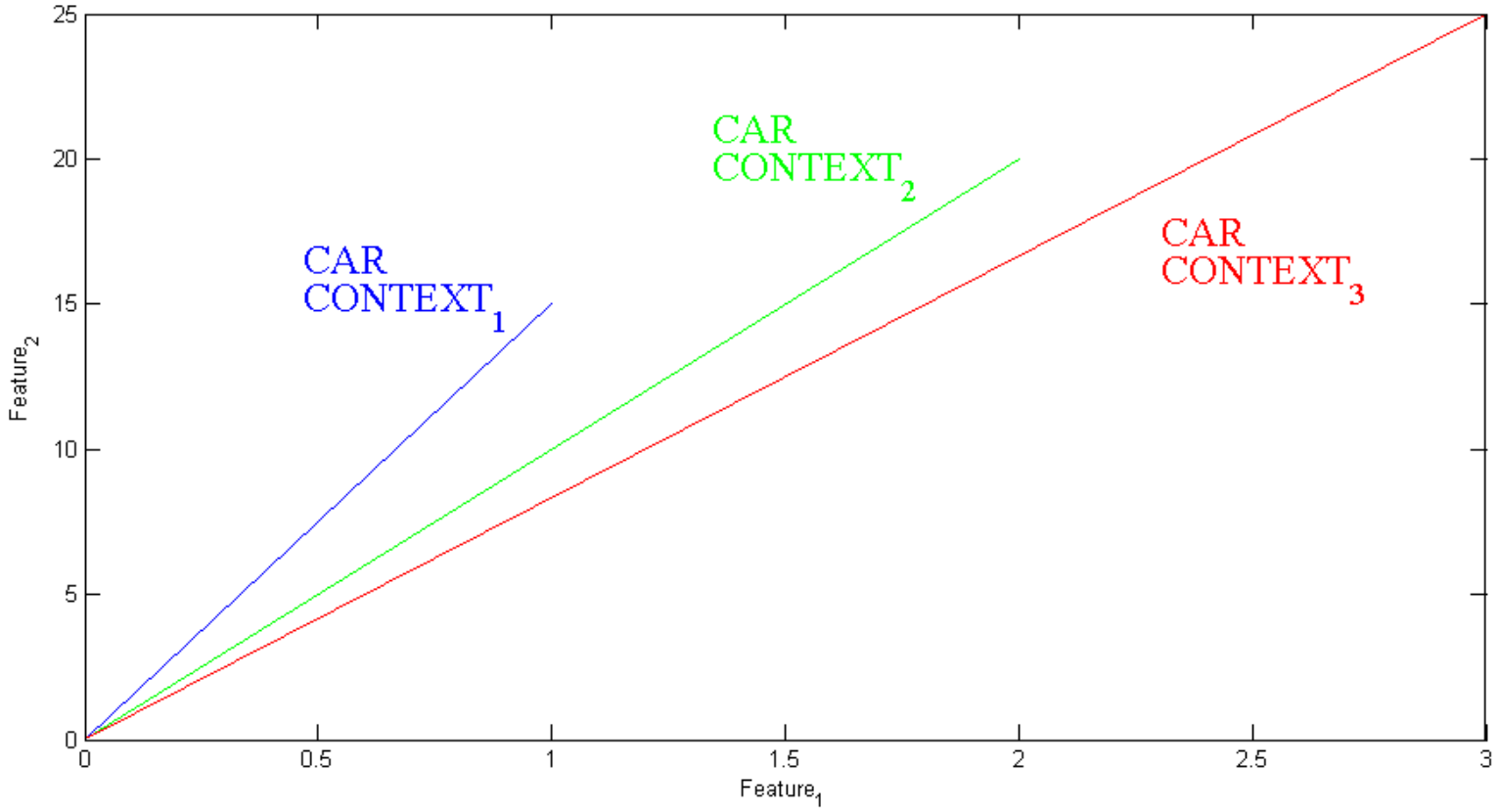
1. Gather its occurrences in a corpus
2. For each occurrence
   1. Select its neighbouring context words
   2. Replace each context word by a precomposed co-occurrence vector
   3. Add these co-occurrence vectors together to get its context vector
3. Calculate similarities between these context vectors and take the average

Vector addition

I ate McDonald's in my car

$\text{contextvector} = \text{vector}_{\text{EAT}} + \text{vector}_{\text{MCDONALD'S}} + \ldots$

Individual term occurrences



To be filled in

Of course our methodology leaves open the choice of a number of parameters:

- features composing the word vectors
- weighting of these features
- window making up our 'context'
- distance metric
- ...

Reference

Peirsman, Y. Crossing Corpora. Modelling Semantic Similarity across Languages and Lects. Ph.D. diss. KU Leuven. 2010.

Sagi, E., Kaufmann, S. & Clark B. (2009a). Semantic density analysis: Comparing word meaning across time and phonetic space.

Overview

- Research question
- From informativeness to corpus linguistics
- Methodology
- Case study

Corpus

Leuvens Nieuws Corpus (LeNC)

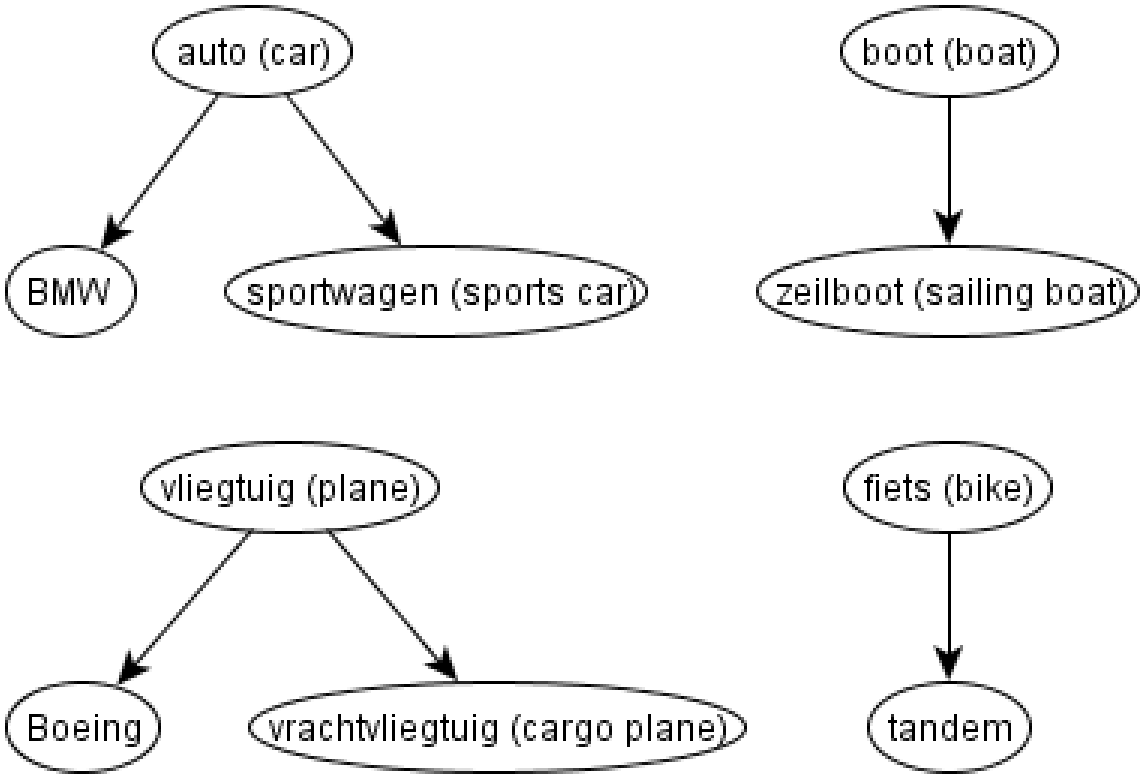
- consists of 6 Flemish newspapers from the period 1999-2005
- totals roughly 750 million words
- syntactically parsed by Alpino parser

Concept choice

We had a look at different means of transportation. We required appearance in either the dictionary (Van Dale) or (Dutch) Wikipedia for each of our concepts.

Concept choice

Concept choice



Concept choice

- Lots of concepts found in the subconcept domain

	# SUBCONCEPTS
CAR	342
BOAT	107
BIKE	34
PLANE	89
ALL	572

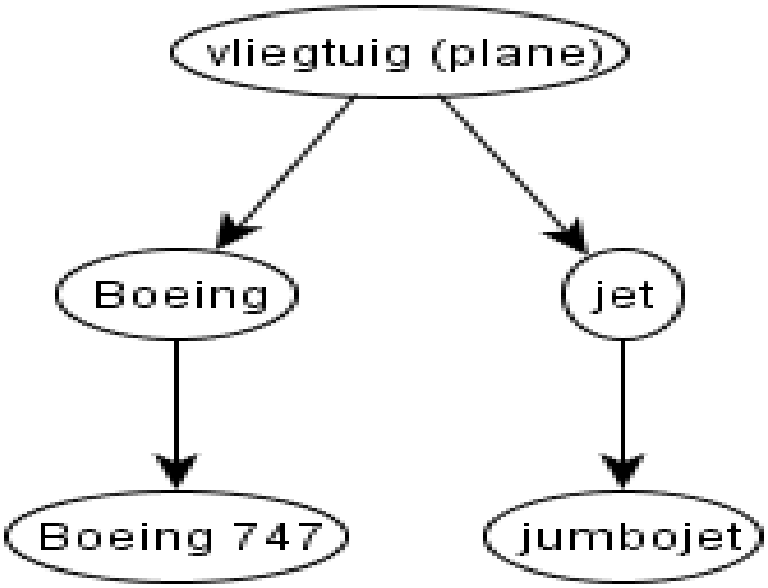
Concept choice

- Lots of occurrences per concept

	# OCCURRENCES		# OCCURRENCES
CAR	569945	SUB-CAR	320526
BOAT	71916	SUB-BOAT	56670
BIKE	102788	SUB-BIKE	15136
PLANE	60715	SUB-PLANE	40409
BASIC LEVEL	805364	SUBORDINATE	432741
ALL	1238105		

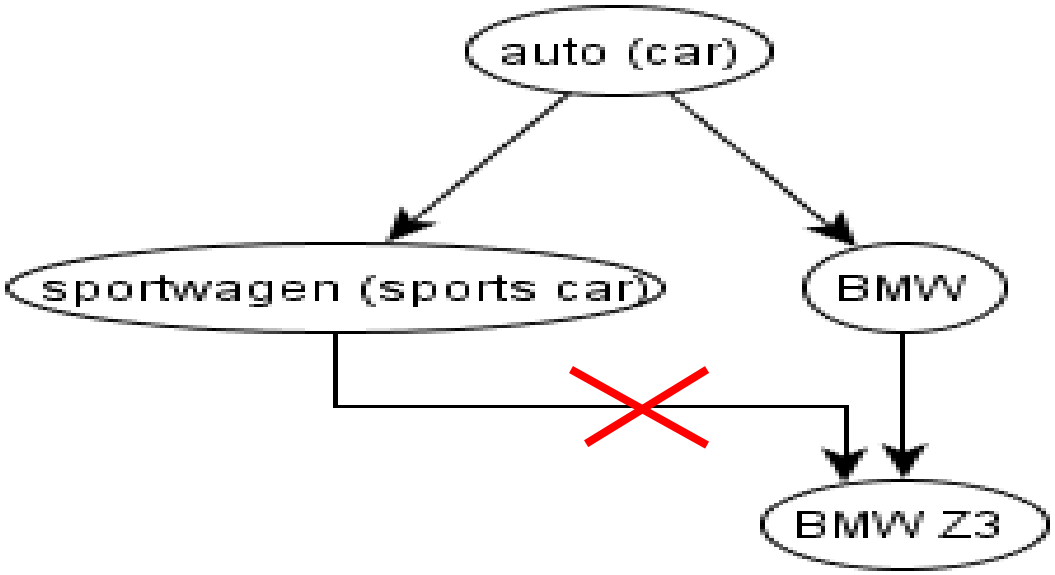
Taxonomy

- Conservativeness in extra subclassing: only where lexical analysis suggests so

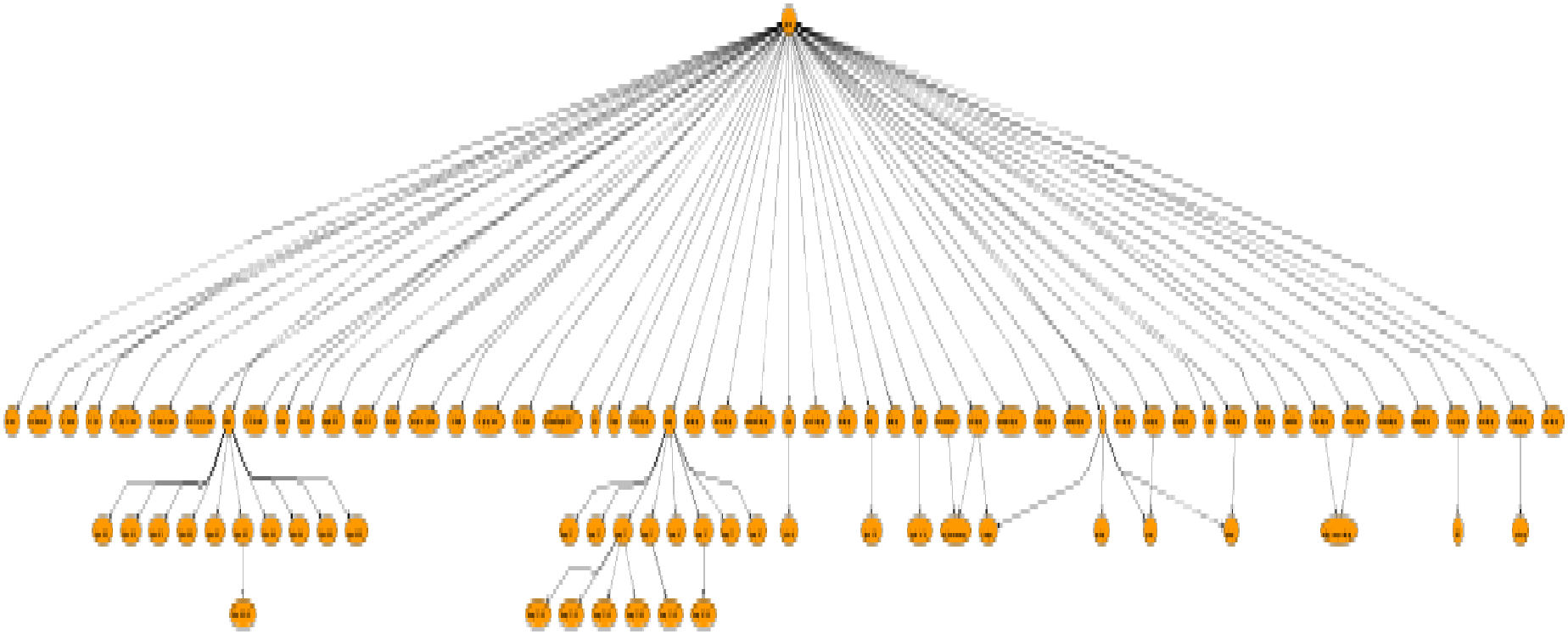


Taxonomy

- Conservativeness in extra subclassing: only where lexical analysis suggests so



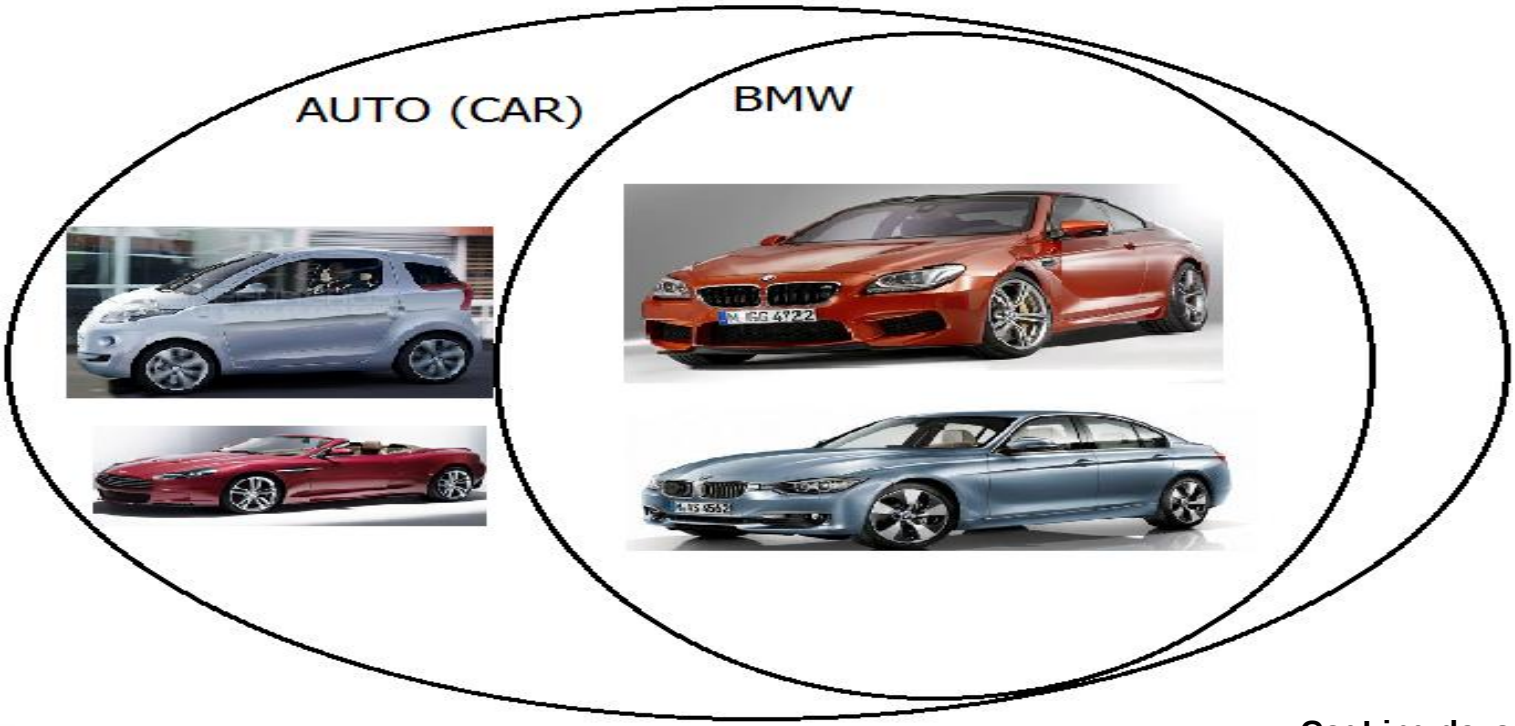
Taxonomy



Hypothesis

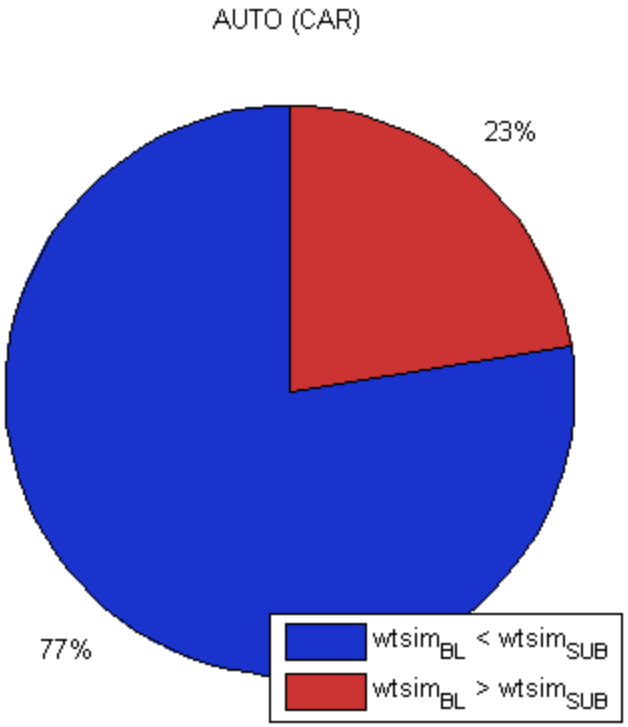
within-term similarity$_{\text{BASIC LEVEL TERM}}$

< within-term similarity$_{\text{SUBORDINATE TERM}}$



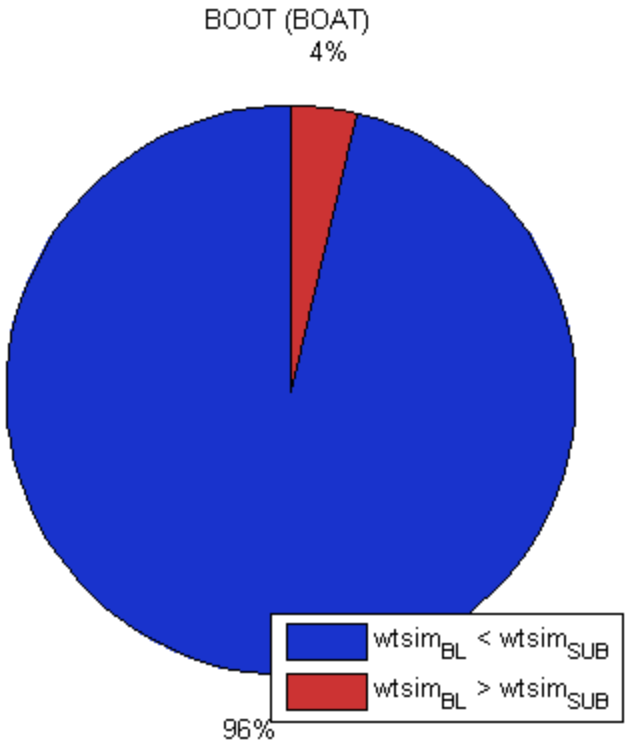
Results

342 comparisons



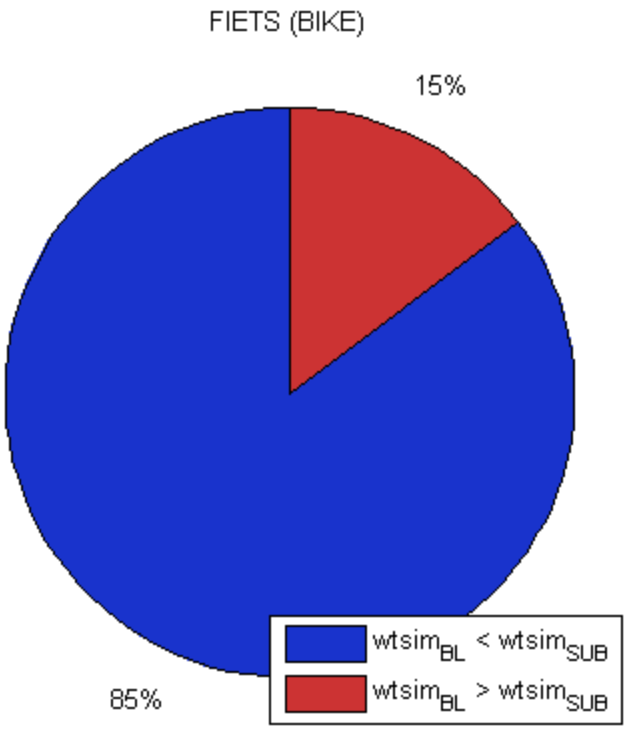
Results

107 comparisons



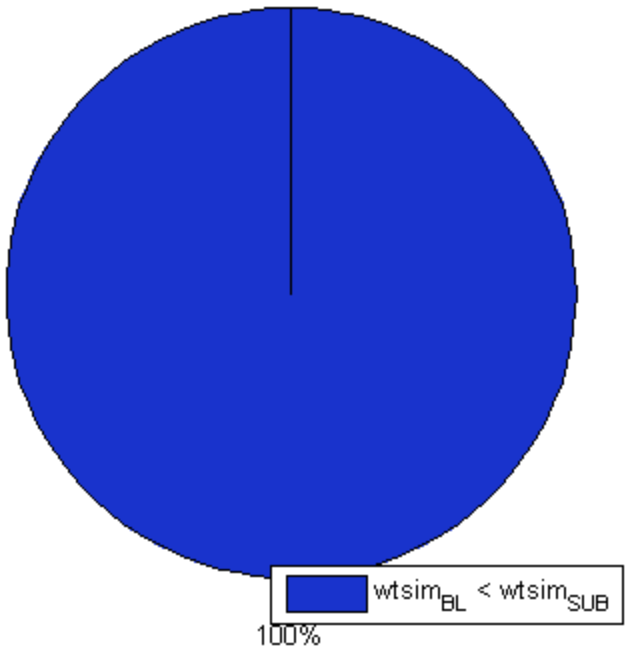
Results

34 comparisons



Results

89 comparisons



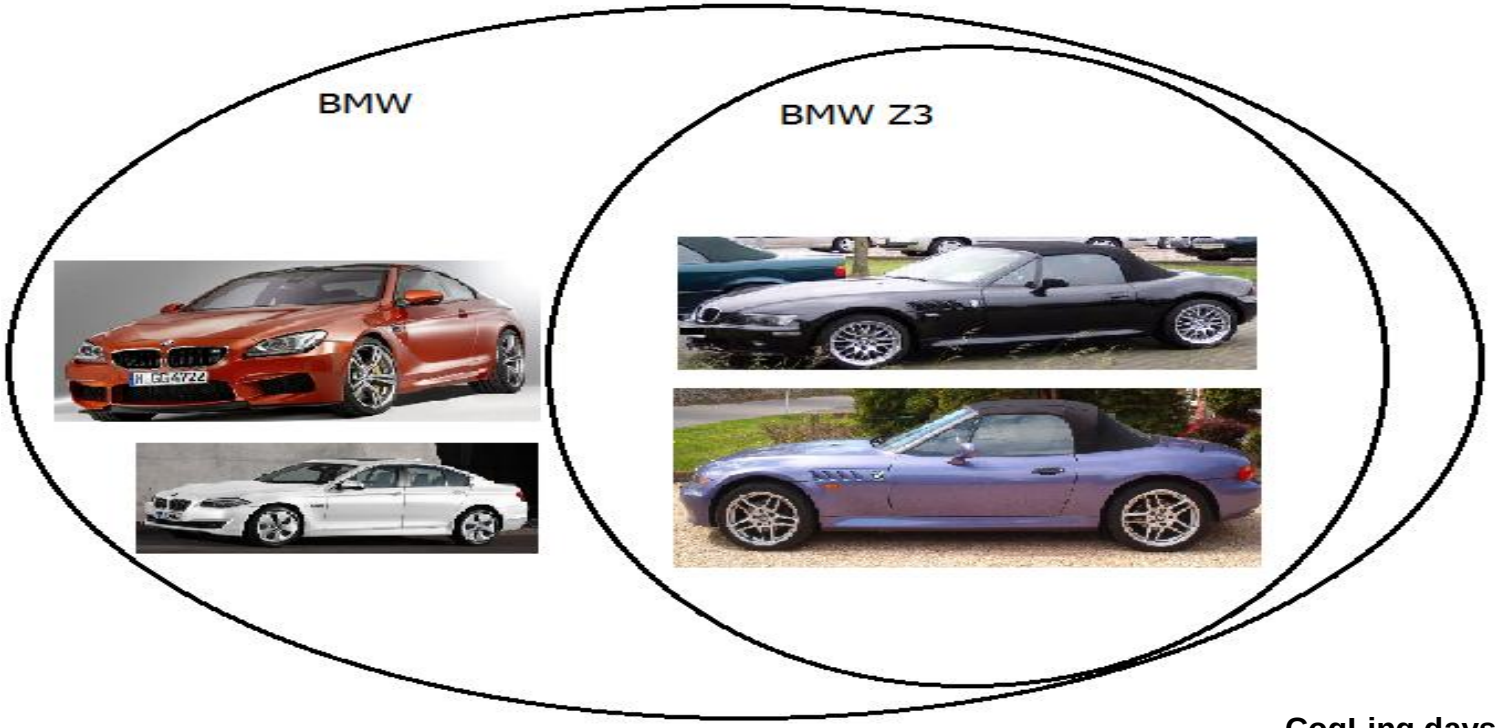
Results

	nb comparisons	succes (%)
auto (car)	342	77
boot (boat)	107	96
fiets (bike)	34	85
vliegtuig (plane)	89	100
	572	85

Hypothesis

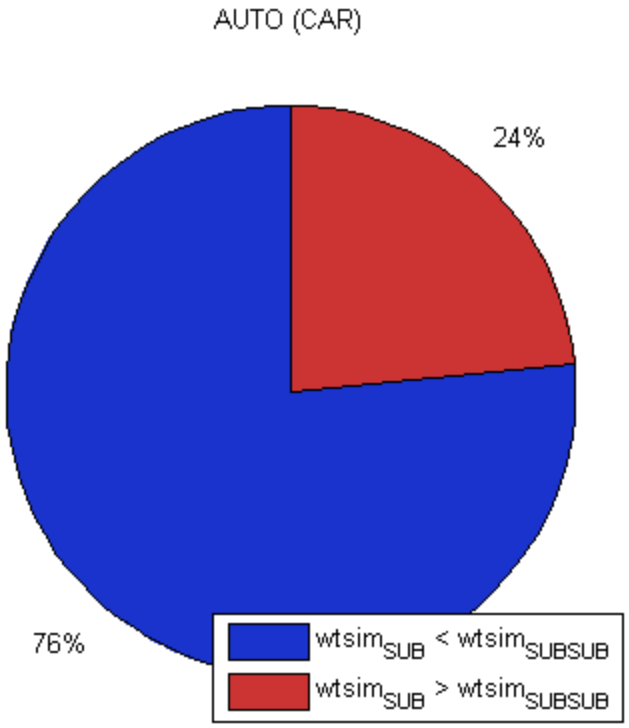
within-term similarity$_{\text{SUBORDINATE TERM}}$

< within-term similarity$_{\text{SUB-SUBORDINATE TERM}}$



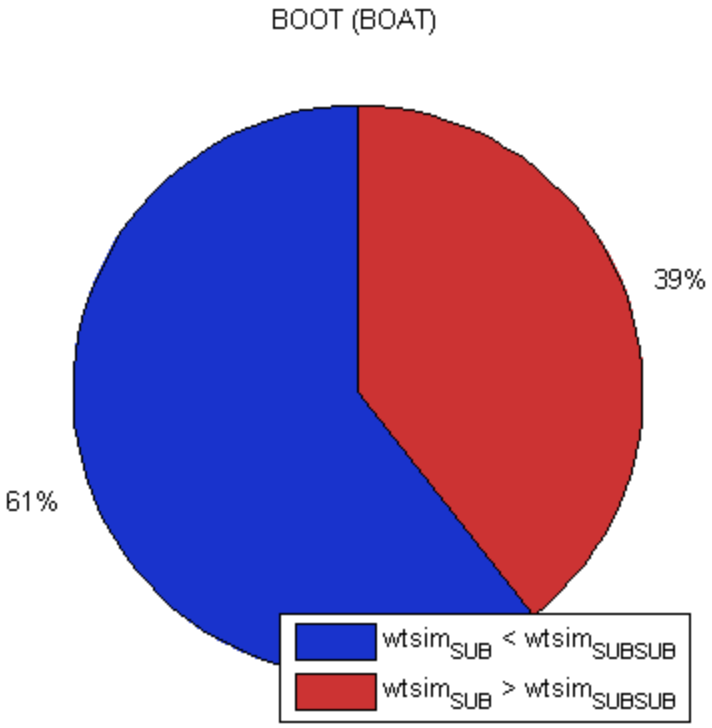
Results

221 comparisons



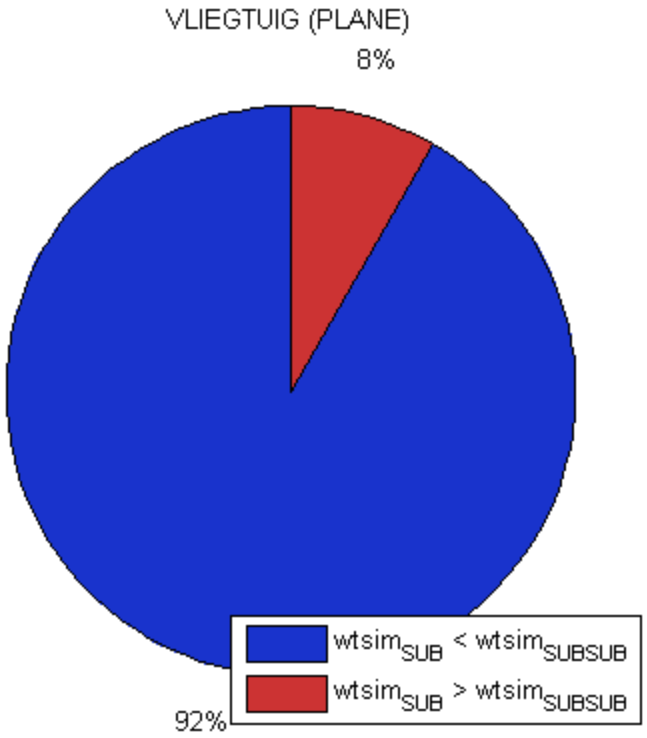
Results

33 comparisons



Results

48 comparisons



Results

	nb comparisons	succes (%)
auto (car)	221	76
boot (boat)	33	61
vliegtuig (plane)	48	92
	302	77

Conclusion

Going down in the taxonomy, we can indeed observe

- within-term similarity$_{\text{BASIC LEVEL TERM}}$ < within-term similarity$_{\text{SUBORDINATE TERM}}$
- within-term similarity$_{\text{SUBORDINATE TERM}}$ < within-term similarity$_{\text{SUB-SUBORDINATE TERM}}$

Conclusion

We should however stress the fact these findings really concern basic level concepts and what's beyond. One cannot simply extrapolate these conclusions to superconcepts !

for further information:

http://wwwling.arts.kuleuven.be/qlvl

stijn.storms@arts.kuleuven.be