Cornetto and WordNet: network of relations based on the words of a language

Isa Maks

e.maks@vu.nl

Computational Lexicology & Terminology Lab
Faculty of Arts
VU University Amsterdam

Cornetto

- STEVIN 2005-2008

Combination of two existing lexical resources (Dutch WordNet and Reference Resource Dutch)

- CLARIN 2012

Standardized to LMF (lexicon markup framework)

ISO data categories

demo

Cornetto

A lexical resource for Dutch which combines

- Syntagmatic information
- Lexical units with detailed information on specific word meanings ranging from form, morphology, syntax, case frames, combinatorics, semantics to pragmatics
- Paradigmatic information
- Synsets built from synonymous lexical units and related through semantic relations

about words and their senses

Overview of the presentation

- Cornetto
  - Lexical Units
  - Synsets and wordnets
  - Links to other resources
- Applications
- Availability

syntagmatic vs. paradigmatic

John	**bought purchased got acquired**	**a(n)**	**expensive costly pricey**	**old new**	**bicycle motorcycle car**

A syntagmatic relationship refers to the relationship a word has with other words that surround it: they are about the lexical company the word keeps.

A paradigmatic relationship refers to the relationship between words that are the same parts-of-speech and which can be substituted for each other in the same position within a given sentence: these words can be substituted.

Together they constitute the identity of an item

Lexical Units

- Words are alphabetically ordered
- And divided in senses or lexical units

opscheppen

opscheppen-v-1

- Definition: m
  over praten (t

opscheppen-v-2

- Definition: v

toe | tō |

noun

1 any of the five digits at the end of the human foot: *he cut his big toe on a sharp stone*.
- any of the digits of the foot of a quadruped or bird.
- the part of an item of footwear that covers a person's toes.

2 the lower end, tip, or point of something, in particular:
- the tip of the head of a golf club, furthest from the shaft.
- the foot or base of a cliff, slope, or embankment.
- a flattish portion at the foot of an otherwise steep curve on a graph.
- a section of a rhizome or similar fleshy root from which a new plant may be propagated.

Lexical Units

- Lemma and word forms
- Morphology and morpho-syntax
- Syntax, syntactic behaviour and syntactic subcategorisation
- Sense and semantics
- Combinatorics

Lemma and wordforms

Plural forms : huis , huizen; (house , houses)

Comparative and superlative forms : groot, groter, grootst ; (large, larger, largest)

Past tense and past participle :deed , gedaan (did done)

Related forms (spelling and form variants): bureau vs. buro organisation vs. organization

Morphology and morpho-syntax

Separability (verbs) : hij **wast** de borden **af** ; he cleans the dishes

Gender (nouns): m, f, n

Adverbial usage (adjectives): slecht (bad, badly)

Syntax and syntactic behaviour

Complementation patterns

ofclause de vraag of … (the question whether)

prep … (opscheppen over , boast about)

syntactic subcategorization frame

hij schept op over haar

PP (constituent) as prepositional object (function)

hij schept (het eten) op

NP as direct object

Sense and semantics

Short definition or resume

Semantic type

Artefact een fabriek bouwen (to build a factory)

Place achter de fabriek wonen (to live behind the factory)

Institution de fabriek is gesloten (the factory is closed)

Syntax and syntactic behaviour

Complementation patterns

ofclause de vraag of … (the question whether)

prep … (opscheppen over , boast about)

syntactic subcategorization frame

hij schept op over haar

PP (constituent) as prepositional object (function)

hij schept (het eten) op

NP as direct object

Combinatorics

Illustrative examples

We saw five monkeys in the zoo

Lexical collocations

Wash the dishes

Grammatical collocations

In terms of

Idioms

Kick the bucket

DEMO

Paradigmatic relations

WordNets and Synsets

Relational model of meaning

animal

kitten

man

boy

cat

tail

dog

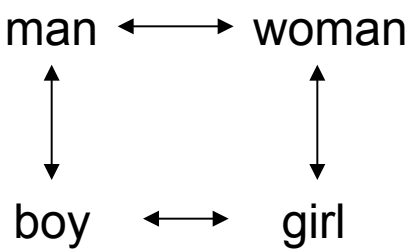
puppy

woman

Relational model of meaning

animal

kitten



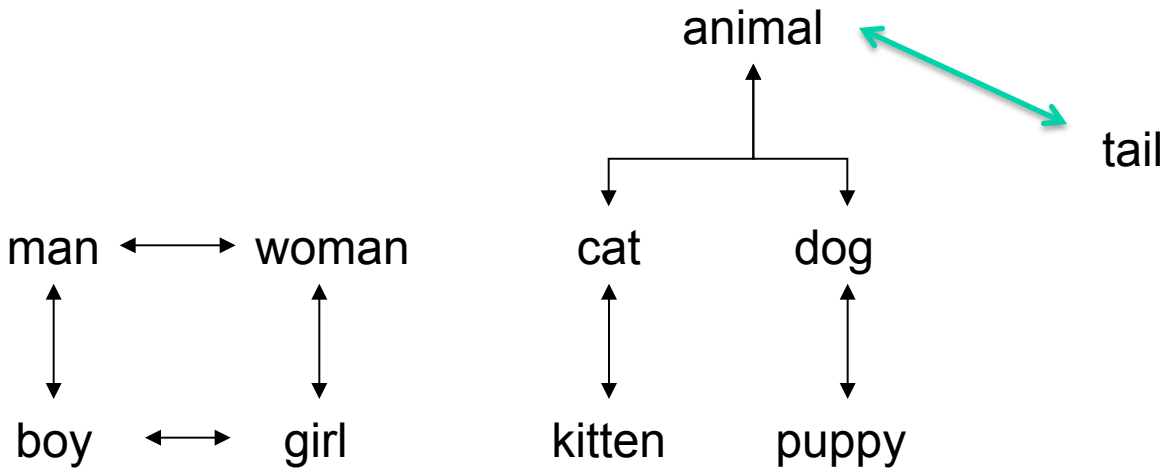
cat

dog

puppy

tail

Paradigmatic information: Relational model of meaning



concepts

Knowledge about concepts

Minimal knowledge about concepts:

- X is a kind of Y
- X has part Y
- an X Ys
- X is Y/has property Y

Most important relation: synonymy

WordNet groups (roughly) synonymous, denotationally equivalent, words into unordered sets of synonyms (synsets)

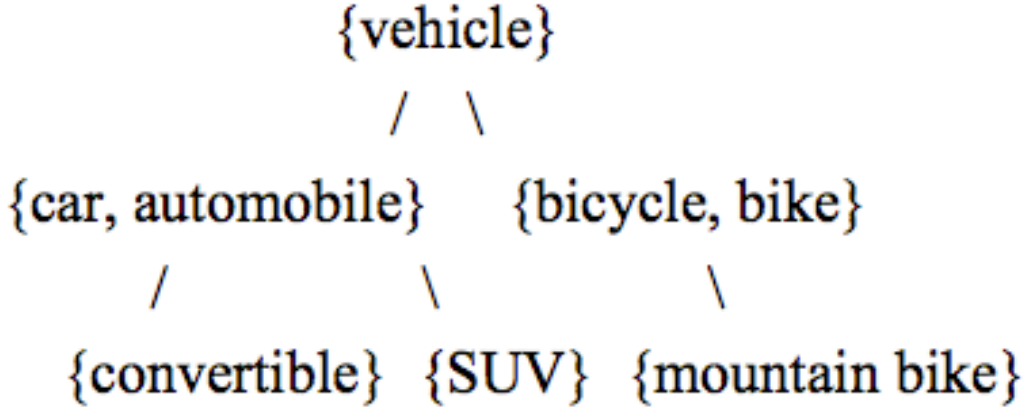
{hit, beat}	{slaan, meppen}
{close, shut}	{sluiten, dichtdoen}
{big, large}	{groot, fors}
{queue, line}	{rij, wachtrij}

Each synset expresses a distinct concept

Most synsets contain > 1 member

HYPONOMY/ HYPERONOMY

Denotes more/less general concepts:



"A car is is a kind of vehicle"

"The class of vehicles contains cars, SUVs, mountain bikes"

Another relation: antonomy

"clang" association between pairs of antonymous words

{big-little}

{wet-dry}

{big, large}

{groot-klein}

{droog-nat, vochtig}

{groot, fors}

{sluiten-openen}

{dood-leven}

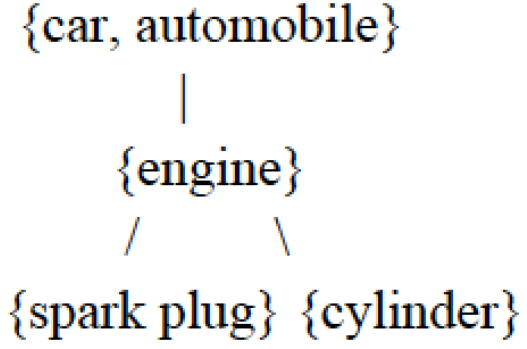
Conceptual-semantic relations

Synsets, the nodes of the network are interrelated via conceptual-semantic relations

- Traditional lexicographer's relations
- Relations from psycholinguistic evidence (word association norms) : the mental lexicon

Meronymy among noun synsets

Meronymy/holonymy (part/whole)



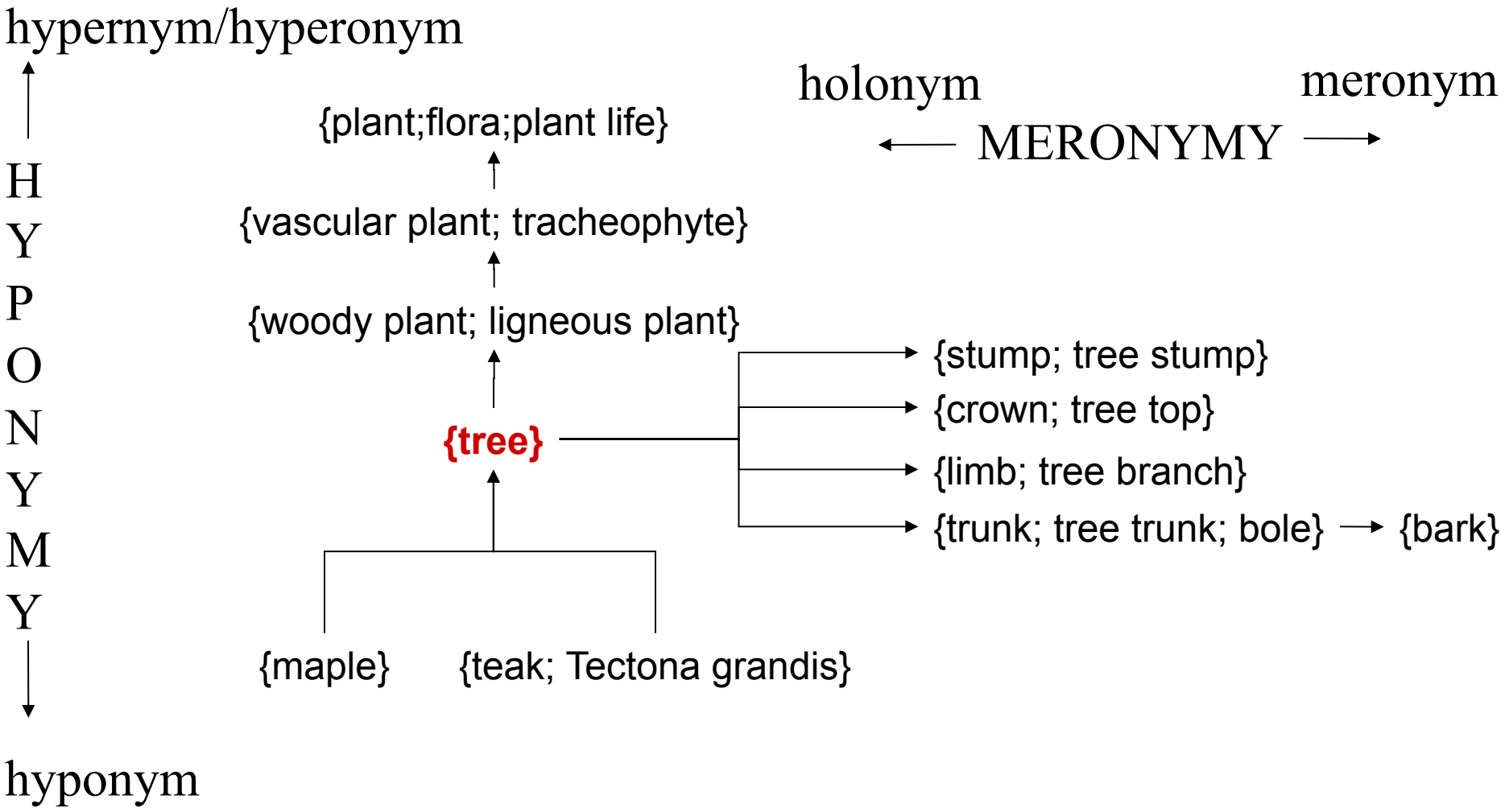
"An engine has spark plugs"
"Spark plus and cylinders are parts of an engine"

Role relations

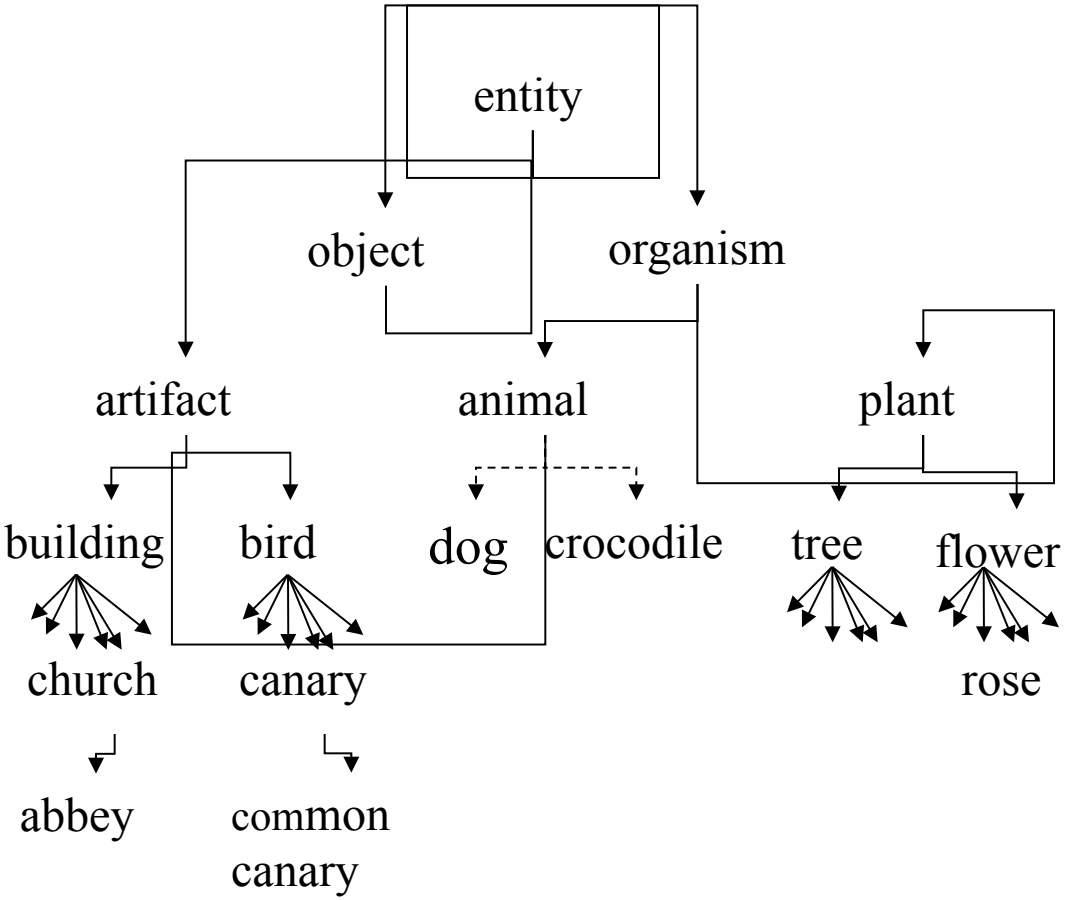
{knife}	ROLE_INSTRUMENT	{to cut}	
{to cut}	INVOLVED_INSTRUMENT	{knife}	*reversed*
{school}	ROLE_LOCATION	{to teach}	
{to teach}	INVOLVED_LOCATION	{school}	*reversed*

Useful when hyponymy does not clarify the position of the concept network, but the word is still closely related to another word.

Wordnet terminology



Lexicalization patterns



top-layer

25 unique beginners

Basic Level Concepts

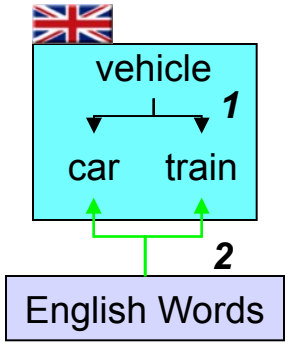
DEMO

What kind of resource is the WordNet?

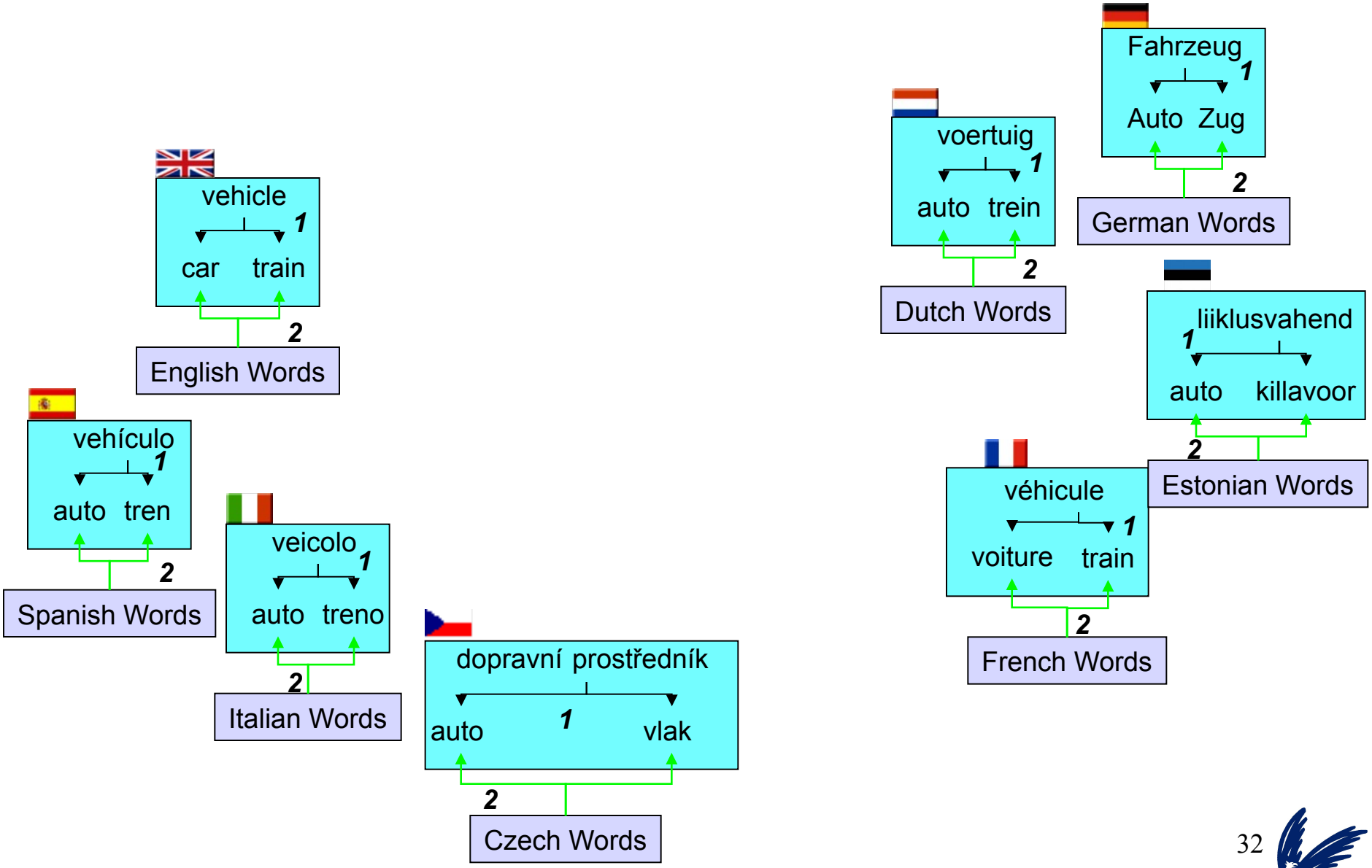
- **http://wordnet.princeton.edu/**
- Developed by George Miller and his team at Princeton University, as the implementation of a mental model of the lexicon
- Mostly used database in language technology
- Enormous impact in language technology development
- Large
- Free and downloadable
- English
- Links to many other knowledge bases

Links to other repositories

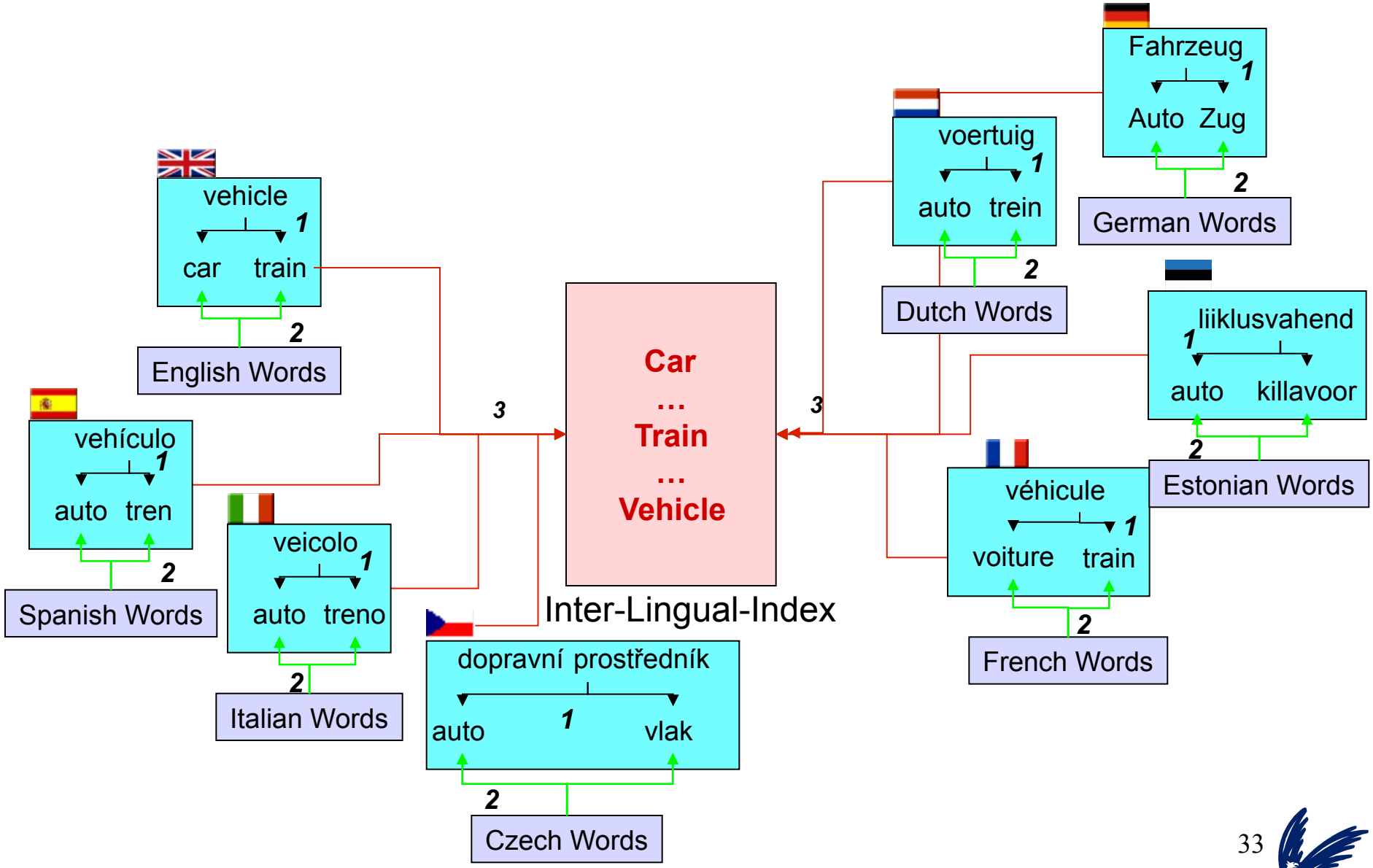
EuroWordNet Model



EuroWordNet Model



EuroWordNet Model



Typical gaps in the English wordnet

- Dutch:

  doodschoppen (to kick to death):

  eq_hyperonym {kill}V and to {kick}V

  aardig (Adjective, to like)

  eq_near_synonym {*like*}V

  cassière (female cashier)

  eq_hyperonym {*cashier}* and to *{woman}*

  kunstproduct (artifact substance)

  eq_hyperonym {artifact} and to {product}

- Spanish:

  alevín (young fish):

  eq_hyperonym {fish} and eq_be_in_state {young}

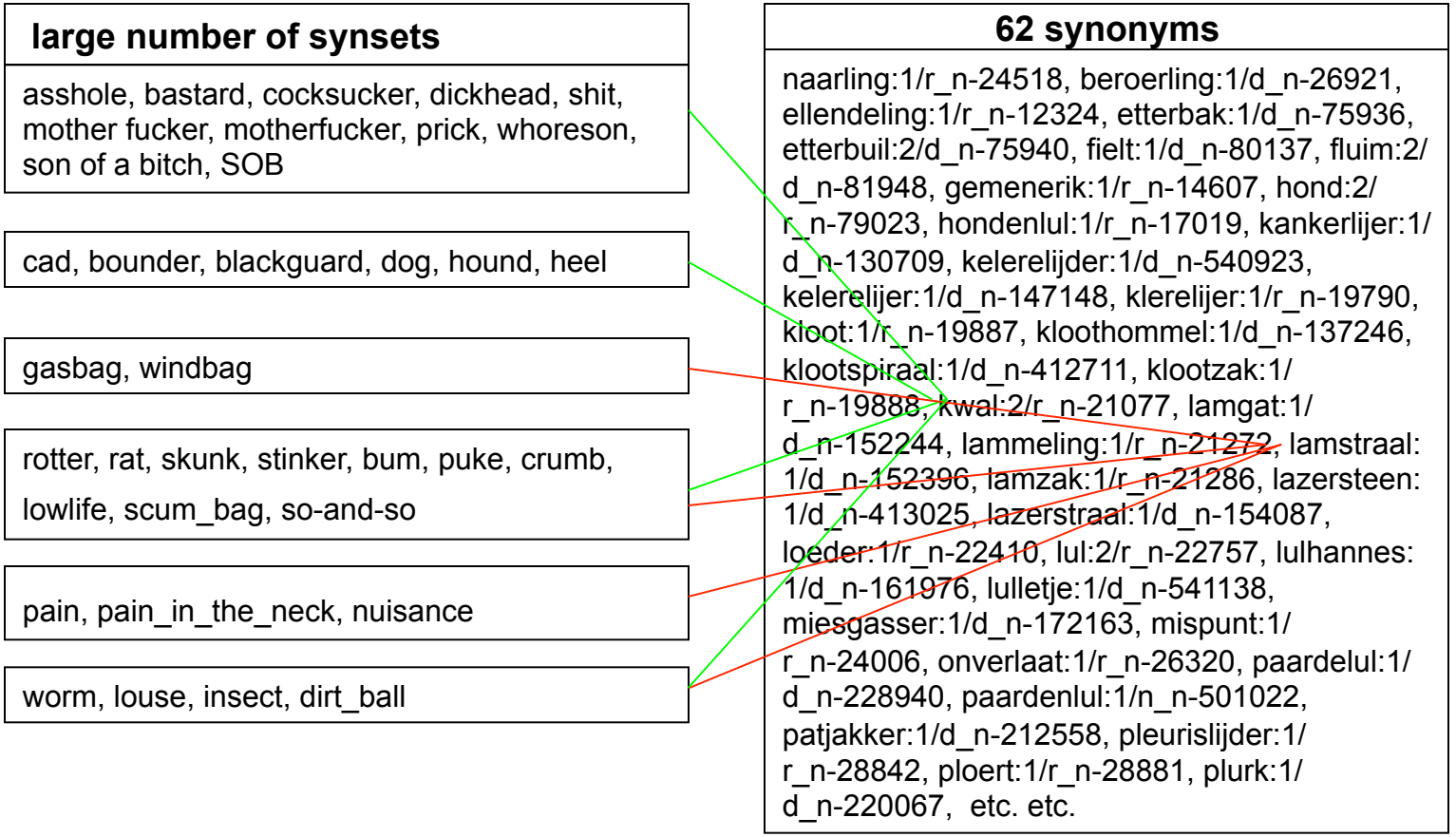
  cajera (female cashier)

  eq_hyperonym {*cashier}* and to *{woman}*

Differences among wordnets

English Wordnet

Dutch Wordnet



insulting terms for people who are stupid, ridiculous, irritating, lazy, slow, ……

Wordnet versus ontology

- **Wordnet**:
  - autonomous **language-specific** and **complex** lexicalization pattern in a relational network.
  - **Usage**: to predict **substitution** in text for information retrieval, text understanding, text generation, machine translation.
- **Ontologies**:
  - data structure with formally defined concepts.
  - **Usage**: making semantic inferences.

Ontology and lexicon

- Hierarchy of disjunct types:

  Canine => PoodleDog; NewfoundlandDog;
  GermanShepherdDog; Husky

- Lexicon:
  - NAMES for TYPES:

    {**poodle**}EN, {**poedel**}NL, {**pudoru**}JP

    => ((instance x Poodle)
  - LABELS for ROLES:

    {**watchdog**}EN, {**waakhond**}NL, {**banken**}JP

    =>((instance x Canine) and (role x GuardingProcess))

Wordnet to ontology mappings

{teacher}Noun, English
-> sc_domainOf ***human***
-> sc_playRole ***done-by***
-> sc_participantOf ***teach***
{leraar}Noun, Dutch // lit. *male teacher*
-> sc_domainOf ***man***
-> sc_playRole ***done-by***
-> sc_participantOf ***teach***
{lerares}Noun, Dutch // lit. *female teacher*
-> sc_domainOf ***woman***
-> sc_playRole ***done-by***
-> sc_participantOf ***teach***

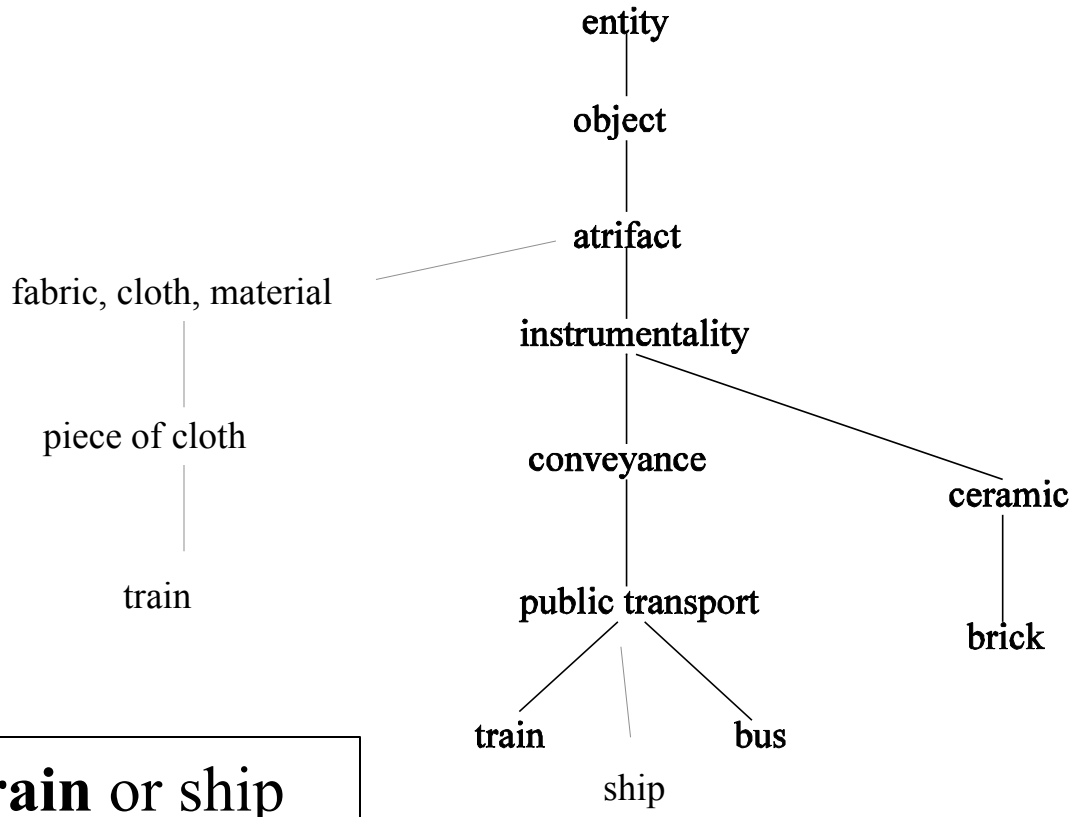
http://www.ontologyportal.org

Usage of Cornetto

Similarity

Sentiment Analysis

Similarity/ Relatedness



It can be delivered by **train** or ship

train

1. public transport provided by a line of railway cars coupled together and drawn by a locomotive
2. piece of cloth forming the long back section of a gown that is drawn along the floor

Generating Sentiment Lexicons

Subjective vs. objective

a burglar alarm (objective)

He cried out in alarm (subjective)

He disapproved of drinking (subjective)

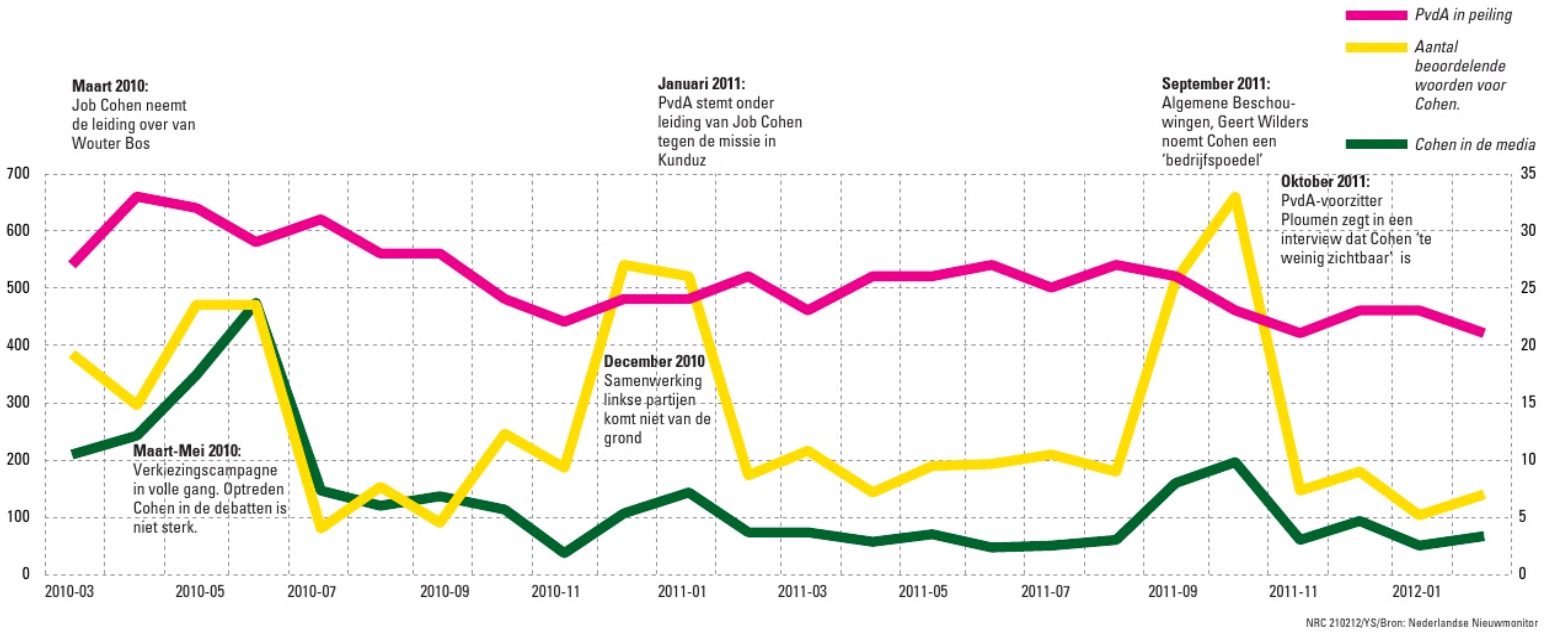
Positive vs. negative

terrible (negative)

fantastic (positive)

review	ur	ann1	ann2
Hotel bezocht ivm jaarlijks weekendje weg met vrienden. Gezellig hotel. Te beginnen met de naam. Hotelbar aanwezig. Persoonlijke en vriendelijke benadering. Wij sliepen op de twee zolderkamers. Niets mis mee. Hotel ligt aan een levendig pleintje met kroegen en restaurants etc. Goede herinneringen. Ontbijt uitstekend.	3	4	4
Ga er zeker niet heen zolang het hotel in renovatie is. De receptie schijnt al klaar te zijn, het is net alsof je aan de incheckbalie bij Schiphol staat en bovendien chagrijnige receptionisten.We wilden een drankje drinken en geen bar te vinden behalve een afschuwelijke ongezellige ruimte. De ober zei dat we beter naar de overkant naar het Holiday Inn konden gaan om even gezellig te zitten. We hebben het ontbijt ook maar gelijk overgeslagen. Prijs/ kwaliteit een 2	-4	-4	-5
Wat een drama hotel, echt vreselijk! Kakkerlakken in de kamer, niet een maar wel 6 en dat in december! 1 handdoekje (schuurpapier) voor 2 personen. Douche delen met 300 andere gasten. Lijkt wel een gevangenis. douche niet schoon, ook al vraag je er om. *wc overgelopen*, *maandverband overal*. belachelijke prijzen voor dit *crematorium*. GEEN AANRADER!P.s. niet via sun air en d reizen boeken. Van onze klachten en beloofd geld nooit iets gezien!	-5	-5	-5
Een Toilet in het midden van de kamer. Een douche zonder cabine. Een ontbijt vol vliegen. Weinig keuze. Lawaai buiten. Geen informatie en lange wachtrijen. Goed bed en netjes maar vanwege allerlei slordigheden, ongeduldig personeel, zeker niet voor herhaling vatbaar	-4	-5	-5
Zoals de andere bezoekers al zeggen. Het hotel is echt prima. Het eten is heerlijk nen op de sevice en de vriendelijkheid is ook niets aan te merken. Wacht alleen nog een paar jaar voor dit deel van Dubai. Het is een enorme bouwput en de sfeer van 1001-nacht is ver te zoeken.	4	-2	3
Goed Hotel, midden in Down Town. Echt een zaken hotel. voor vakantie zou ik iets anders uitzoeken.	4	-2	-2
Het receptie personeel moet cliënten met U aanspreken en niet met 'je'	4	-2	-2

COHEN WEET STEEDS MINDER GOED DE MEDIA-AANDACHT OP TE EISEN



NRC 210212/YS/Bron: Nederlandse Nieuwmonitor

JUIST OP DE BELANGRIJKE MEDIAMOMENTEN LIET COHEN HET AFWETEN

• Was het werkelijk de 'mediawerkelijkheid' die Job Cohen mede de das omdeed, zoals hij gisteren zelf verklaarde? Wetenschappelijk instituut voor de journalistiek De Nederlandse Nieuwsmonitor onderzocht dit samen met de afdeling Communicatiewetenschap van de Vrije Universiteit voor *nrc.next*.

• Cohen heeft moeite met de nieuwe realiteit, zo blijkt. Inmiddels wordt de manier waarop nieuws wordt gebracht allang niet meer door de inhoud, maar door de vorm bepaald. Met de televisie als meest bepalende medium. En precies daarmee had Cohen moeite. Het in één minuut duidelijk maken van een standpunt tijdens een televisiedebat. Het meelachen met Rutger Castricum. Wilders aanpakken als het er echt op aankomt.

• De onderzoekers hebben een inhoudsanalyse uitgevoerd van alle artikelen in *de Volkskrant, Trouw, de Telegraaf, AD* en *NRC Handelsblad* tussen het aantreden van Cohen op 12 maart 2010 en gisteren. Er is gekeken naar de hoeveelheid media-aandacht voor Cohen, maar ook naar de aard van deze berichtgeving. Deze 'sentimentanalyse' kan op basis van trefwoorden bepalen of een artikel positief, neutraal of negatief van aard is. De betrouwbaarheid van deze analysevorm ligt volgens De Nieuwsmonitor tussen de 60 en 70 procent.

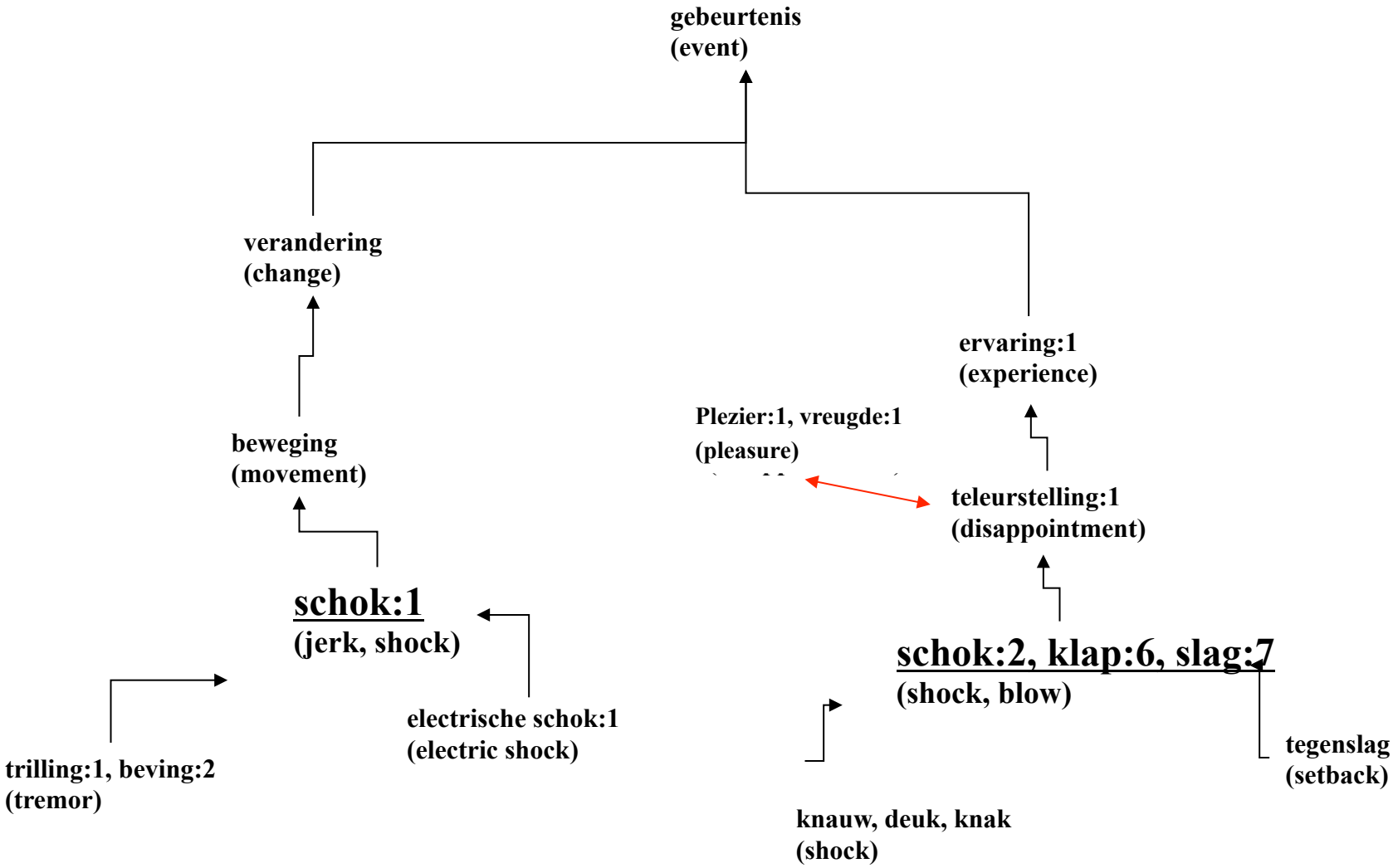
• Als hij in maart 2010 leider van de PvdA wordt, spreken de krantenartikelen vol vertrouwen over Cohen, blijkt uit het onderzoek. Cohen is de gedoodverfde winnaar van de verkiezingen. Dit vertrouwen zakt snel weg. In het begin van de campagne heeft Cohen het zwaar in de debatten. Enkele maanden na zijn aantreden is de teneur van de artikelen al een stuk negatiever.

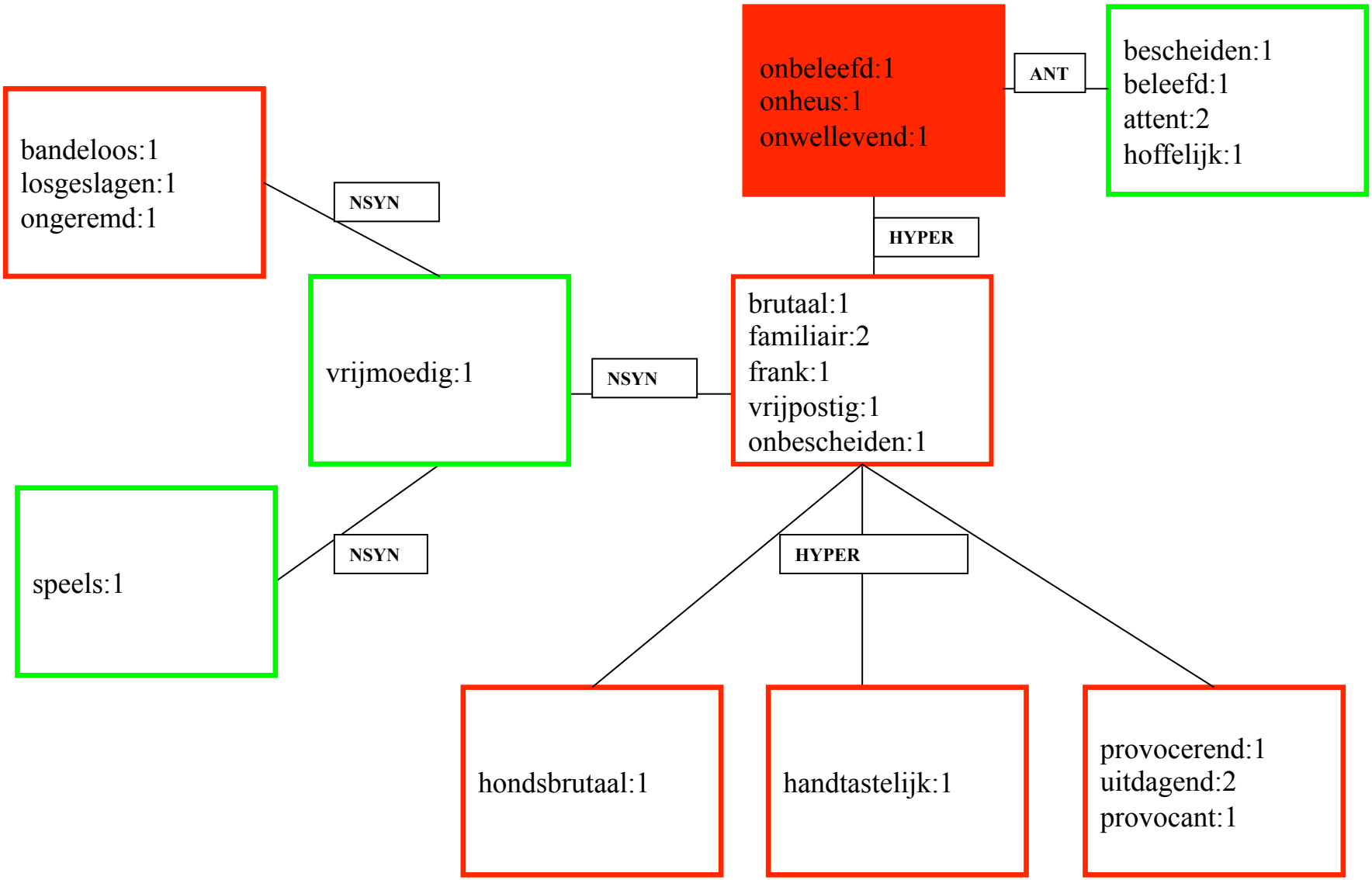
• Daarna zal het niet meer goed komen met de PvdA-leider. Cohen laat het juist op de belangrijke mediamomenten afweten, zo is te zien in de grafiek hierboven. De pieken in de beoordelingslijn laten zien wanneer het er bij Cohen echt om ging: de momenten dat hij het meest werd 'beoordeeld' in de media. En voor Cohen waren dat nu juist elke keer momenten waarop hij de meeste kritiek kreeg: de debatten, Wilders die hem een bedrijfspoedel noemde en PvdA-voorzitter Ploumen die Cohen een gebrek aan zichtbaarheid verweet.

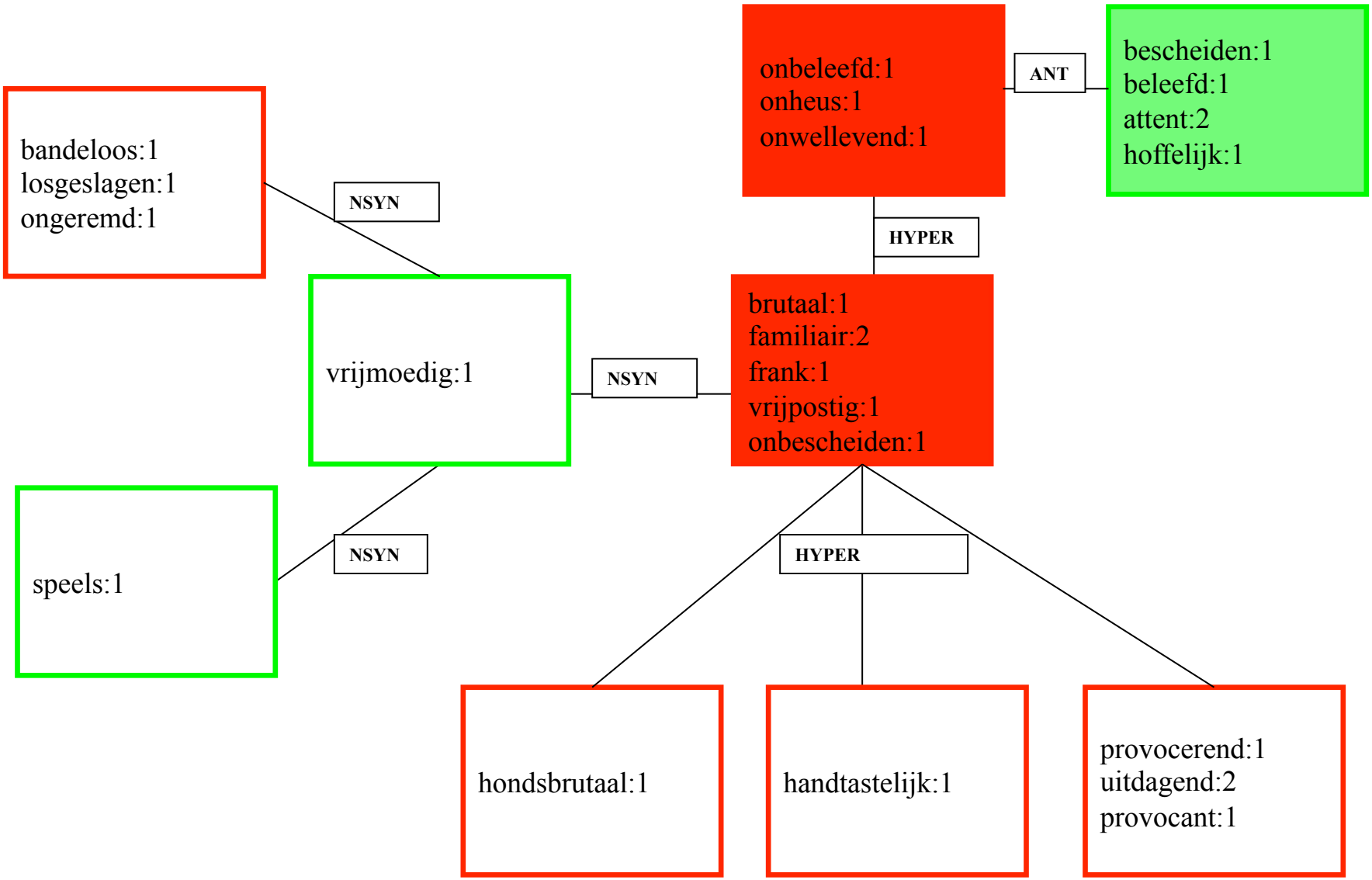
STIJN BRONZWAER
Redacteur Media

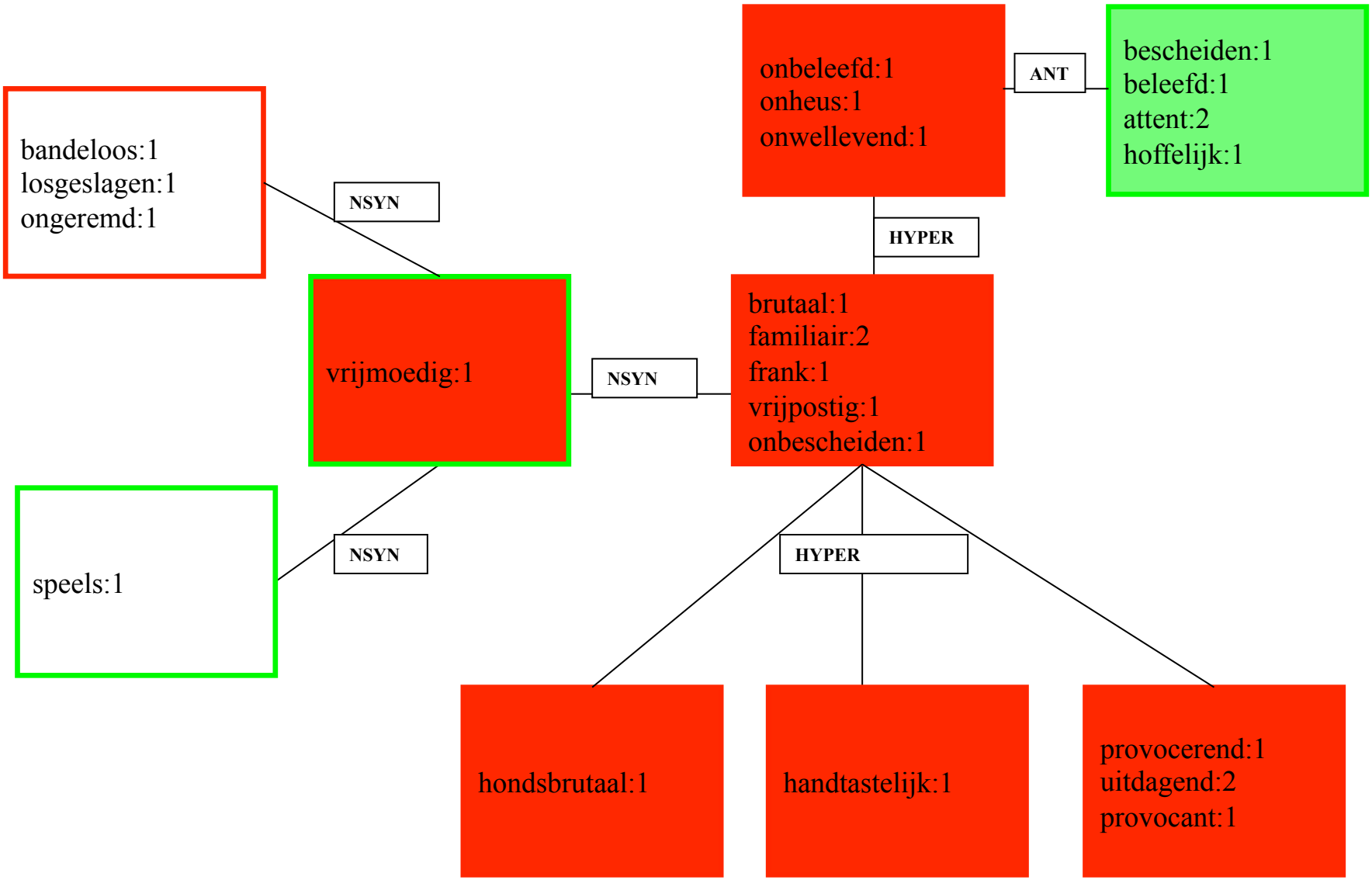
Lees het hele onderzoek op www.nieuwsmonitor.net

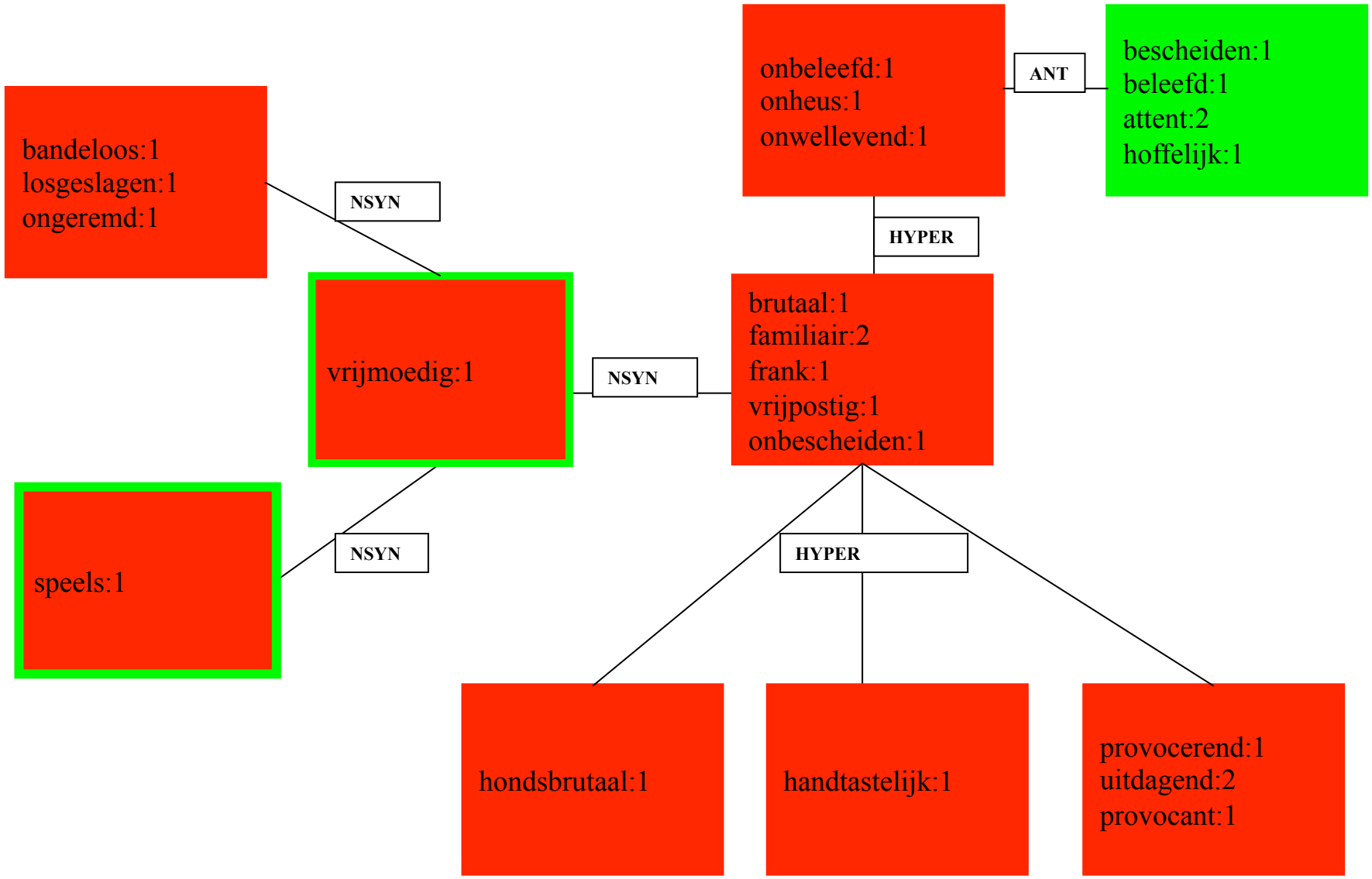
Method I : Wordnet propagation Dutch Wordnet Cornetto




onbeleefd:1
onheus:1
onwellevend:1
ANT
bescheiden:1
beleefd:1
attent:2
hoffelijk:1
bandeloos:1
losgeslagen:1
ongeremd:1
NSYN
HYPER
vrijmoedig:1
NSYN
brutaal:1
familiair:2
frank:1
vrijpostig:1
onbescheiden:1
speels:1
NSYN
HYPER
hondsbrutaal:1
handtastelijk:1
provocerend:1
uitdagend:2
provocant:1


onbeleefd:1
onheus:1
onwellevend:1
ANT
bescheiden:1
beleefd:1
attent:2
hoffelijk:1
bandeloos:1
losgeslagen:1
ongeremd:1
NSYN
HYPER
vrijmoedig:1
NSYN
brutaal:1
familiair:2
frank:1
vrijpostig:1
onbescheiden:1
speels:1
NSYN
HYPER
hondsbrutaal:1
handtastelijk:1
provocerend:1
uitdagend:2
provocant:1


onbeleefd:1
onheus:1
onwellevend:1
ANT
bescheiden:1
beleefd:1
attent:2
hoffelijk:1
bandeloos:1
losgeslagen:1
ongeremd:1
NSYN
HYPER
vrijmoedig:1
NSYN
brutaal:1
familiair:2
frank:1
vrijpostig:1
onbescheiden:1
speels:1
NSYN
HYPER
hondsbrutaal:1
handtastelijk:1
provocerend:1
uitdagend:2
provocant:1


onbeleefd:1
onheus:1
onwellevend:1
ANT
bescheiden:1
beleefd:1
attent:2
hoffelijk:1
bandeloos:1
losgeslagen:1
ongeremd:1
NSYN
HYPER
vrijmoedig:1
NSYN
brutaal:1
familiair:2
frank:1
vrijpostig:1
onbescheiden:1
speels:1
NSYN
HYPER
hondsbrutaal:1
handtastelijk:1
provocerend:1
uitdagend:2
provocant:1

Method II : machine learning with Cornetto features

Lexical unit features

Gradability

awful, more awful, most awful

*Left hand, *lefter hand, *leftest hand*

Attributive vs. predicative use

*The chemical lab vs. *the lab is chemical*

Complementation patterns

Syntactic Complementation (lack of), valency

more syntax = less subjectivity and zero syntax = more subjectivity.

*Hij zeurt. (he nags) vs. *hij zegt. (he says)*

Semantic type

Emotion, cognitive_verb, (subjective words) Action, artifact (objective)

Selectional restrictions (nouns and verbs)

He annoyed me (subject:person/ object:person) vs. he pushed the car (object :non-human)

Definition overlap

Gloss overlap (Lesk): definitions of subjective words have more in common with each er than definitions of objective words

Method II : machine learning with Cornetto features

Synset features

WN similarity: subjective concepts are more similar and/or more related to each other than to objective concepts

Synset size: synsets with subjective concepts have more synonyms on average than other synsets.

WordNet domain features

SUMO features

Method II uses information from both the lexical unit and the synset part of Cornetto to generate a sentiment lexicon

Results

Propagation (positive vs. negative)

goed/a/positive (good)

zeuren/v/negative (nag)

Machine learning (objective vs. subjective)

goed/a/subjective

zeuren/v/subjective

voetballen/v/objective (play football)

		ML	Prop	ML&prop
ML	n	0.77	0.57	
ML	a	0.76	0.77	
ML	v	0.69	0.65	

2.2 Statistics

The current version of Cornetto (version 2.1) includes the following items:

Lexical entries	127,334
Noun Lexical entries	85,368
Verb Lexical entries	16,502
Adjective Lexical entries	15,458
Multiword Lexical Entries	9,397
Lexical Entries with polarity labels	22,202
Sense group relations	11,812
Sense examples	80,512
Lexical collocations	19,166
Grammatical collocations	10,373
Synsets	70,497
Synset relations (internal relations between Dutch synsets)	91,734
Synset equivalent relations (between Dutch and English Princeton WordNet synsets)	84,031
References to SUMO (from synsets to SUMO)	69,610
References to WordNet domains (from synsets to WordNet domains)	93,165

Table 1: statistics of Lexical Resource Cornetto (version 2.1, May 2013)

Availability Cornetto

Lexical Resource CORNETTO		
demo	http://cornetto.inl.nl/	
demo editor	https://debvisdic.let.vu.nl:9002/	username: gast password: gast
download	http://tst-centrale.org/nl/producten/lexica/cornetto/ Xml-LMF or RDF	Free for research (with a license)