

The Cornetto Database: Architecture and User-Scenarios

Piek Vossen, Irion Technologies/Vrije Universiteit van Amsterdam

Katja Hofmann, Maarten de Rijke, Erik Tjong Kim Sang: ISLA, University of Amsterdam

Koen Deschacht: Interdisciplinary Centre for Law and ICT, University of Leuven



nū | STEVIN

Overview

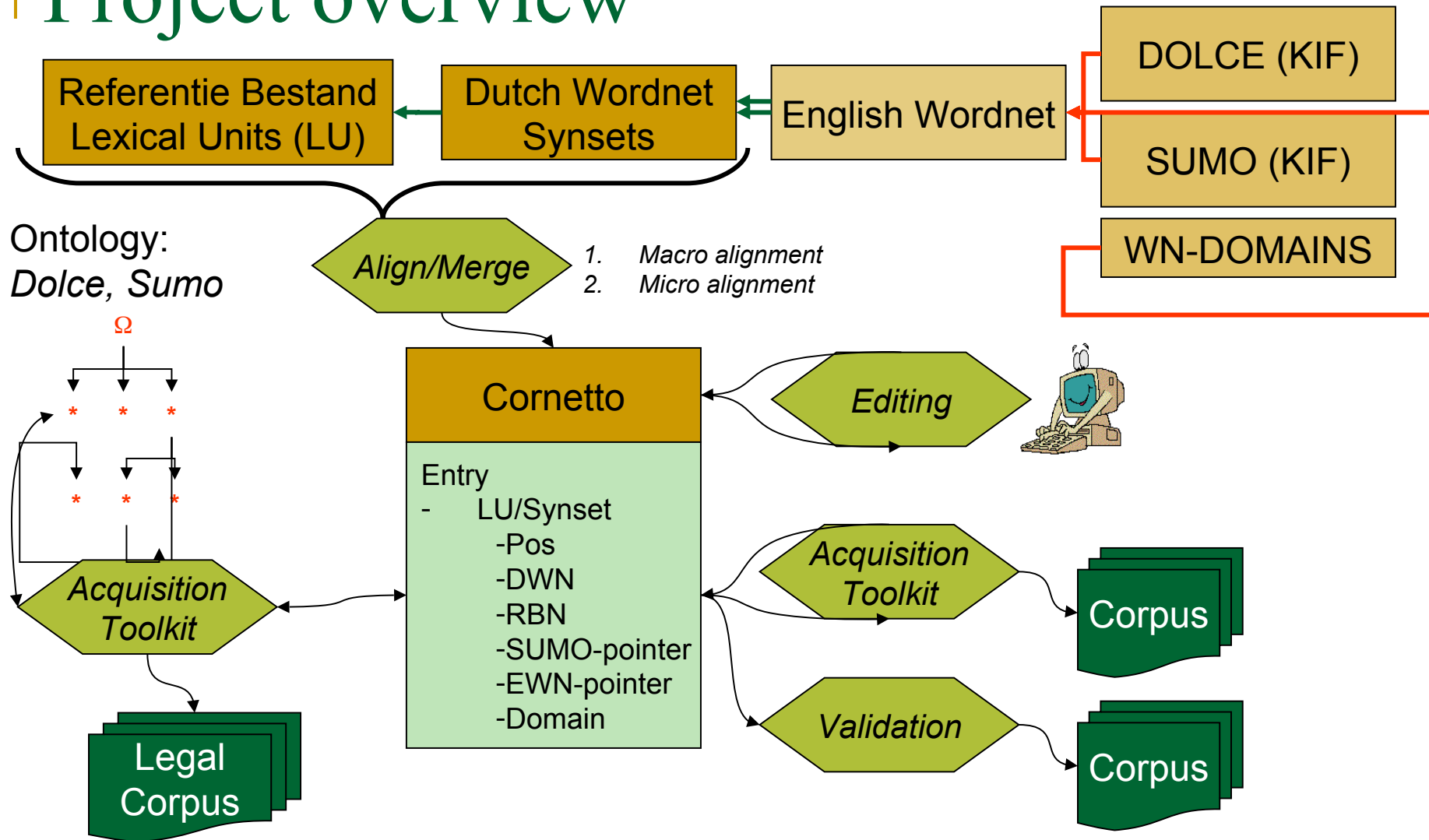
1. Project background information
2. Database design
3. Usage of the database in NLP



Goal of Cornetto

- <http://www.let.vu.nl/onderzoek/projectsites/cornetto/start.htm>
- Stevin tender project to develop a lexical semantic database for Dutch:
 - 40K Entries
 - Generic and central part of the language
- Data:
 - Combination of WordNet and FrameNet
 - Vertical and horizontal semantic relations
 - Combinatorial lexical constraints
 - Aligned with the English Wordnet
 - Extended with an ontology
- Automatic acquisition toolkit
- Started April 2006, ends March 2008
- Licensed from TST-centrale, Nederlandse Taalunie

Project overview



Database design

Lexical Unit & Synsets

- Lexical Unit = form-meaning relation, such that:
 - form = abstract representation of certain realizations;
 - part-of-speech is the same;
 - meaning is the same, where meaning is defined by a reference to a unique Synset;
- Synset = Set of synonyms (LUs) that refer to the same entities in most contexts.
 - Defined by lexical semantic relations;
 - Defined by reference to ontology Terms or KIF expressions involving Terms from the ontology;

Data structure overview

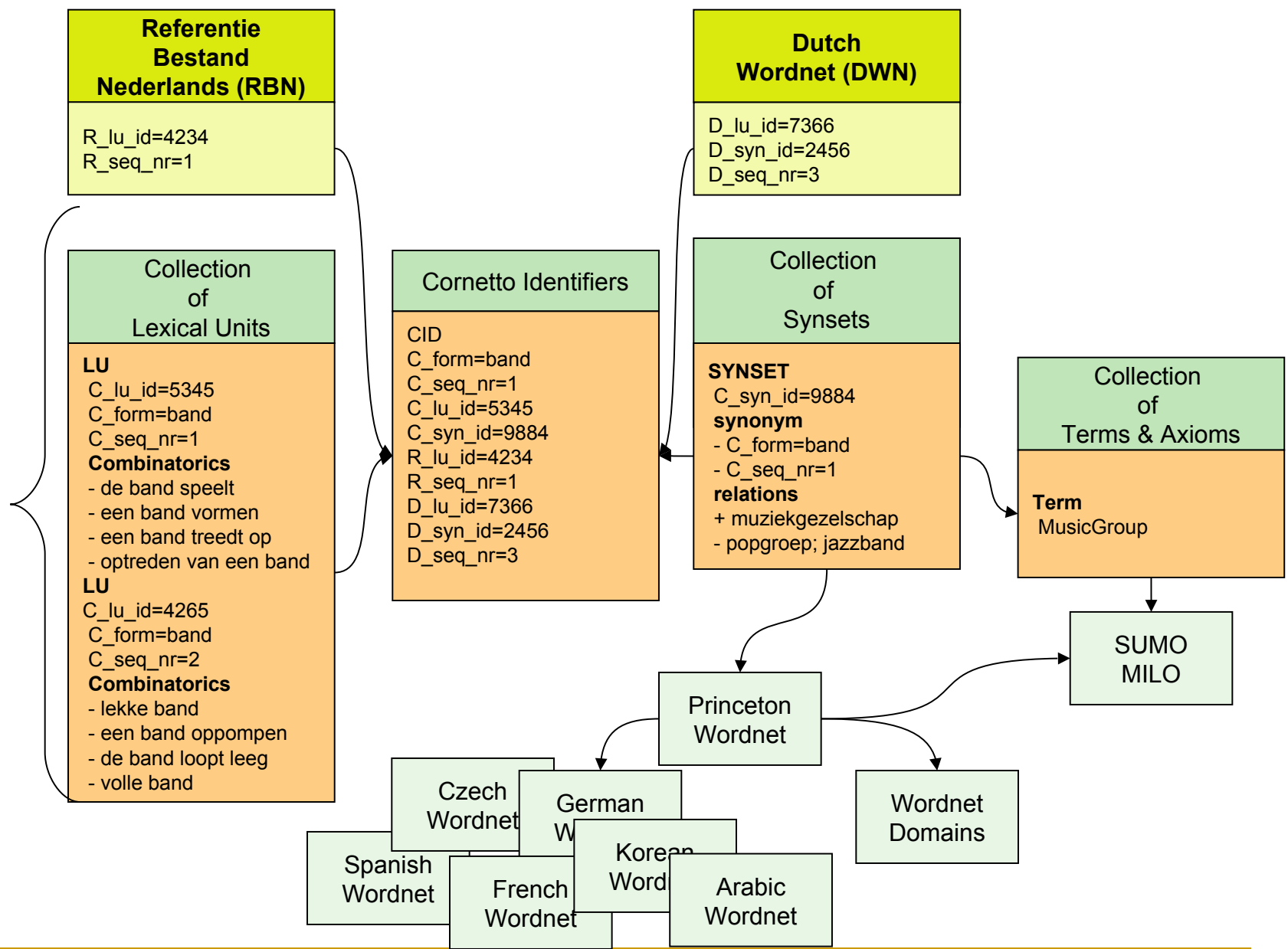
■ Collections:

- ❑ Lexical units (LU): -> mainly derived from RBN
- ❑ Synsets (SY): -> mainly derived from DWN
- ❑ Terms (TE): -> based on SUMO/MILO, linked to PWN
- ❑ Domains (DM): -> based on Wordnet domains

■ Mappings:

- ❑ LU<-> SY
- ❑ SY <-> SY (within Dutch and from Dutch to English)
- ❑ SY <-> TE
- ❑ SY <-> DM

Cornetto Database (CDB)



Combinatoriek

in een band spelen

een band oprichten

de band speelt

Combinatoriek

de band oppompen

een band plakken

een lekke band

de band springt

Combinatoriek

de band starten

op de band opnemen

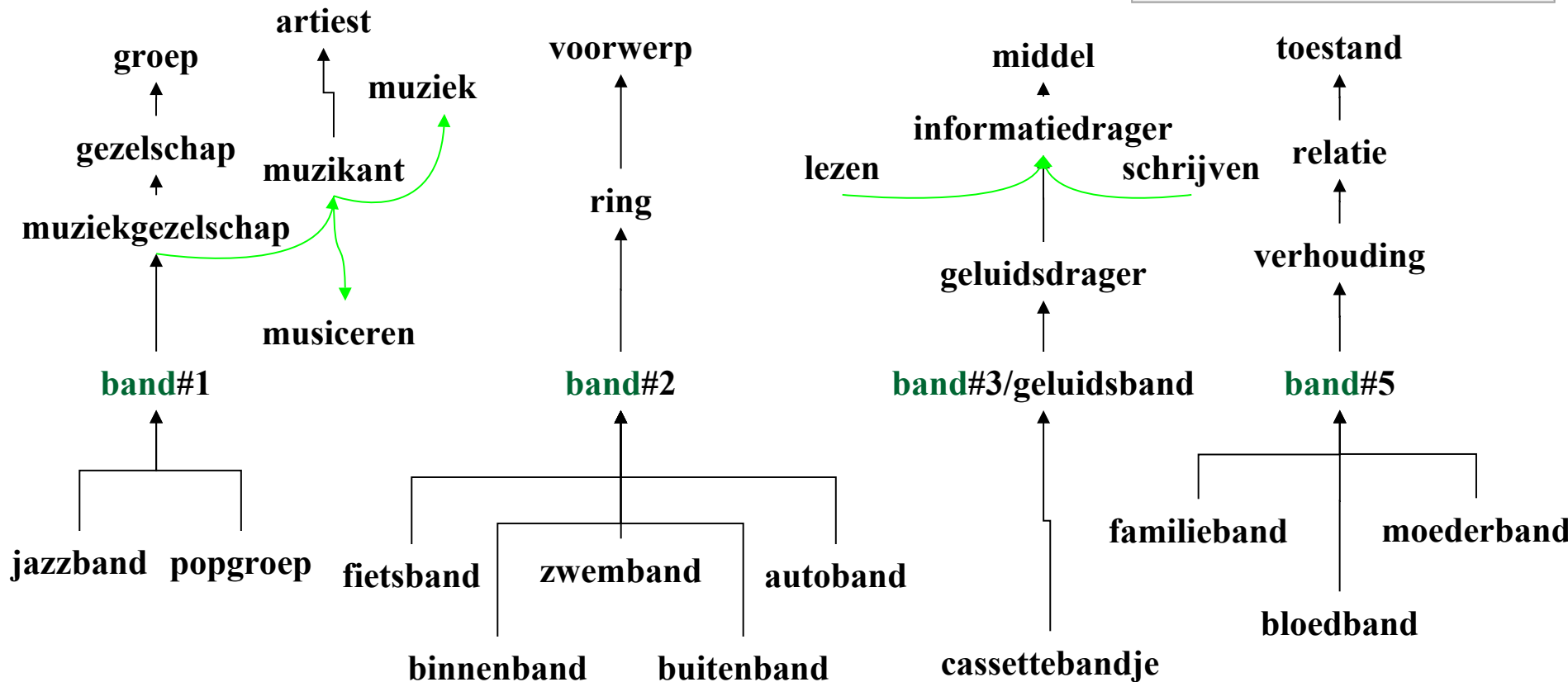
de band afspelen

Combinatoriek

een goede/sterke band

de banden verbreken

een band hebben met iemand



Disambiguating combinations

slag & leveren (fight)
slag & vrije (free style swimming)

punt & delen (share points)
punt & pakken (take a point)
punt & scoren (score a point)
punt & hard (firm point)

trap & geven (to kick)
trap & krijgen (be kicked)
trap & vrij (free kick)
trap & hoge (high stairs)

band & lekke (flat tire)
band & lopende (converter belt)
band & losse (loose relation)
band & spelen (music band plays)
band & starten (start a tape or belt)
band & sterke (strong relation)

bui & flinke (heavy shower)
bui & plaatselijk (local shower)
bui & vallen (there can be shower)
bui & slechte (bad mood)

Prototypical contexts

functie & bekleden (take a position)
functie & hoge (high position)
functie & neerleggen (put down position)
functie & ontheffen (discharge from position)
functie & openen (open up a position)
functie & treden (start a position)
functie & uitoefenen (carry out position)
functie & vervullen (fulfill position)

werk & aannemen (take on work)
werk & bieden (offer work)
werk & gaan & naar (go to work)
werk & half (incomplete work)
werk & staken (strike)
werk & zetten & aan (to put to work)
werk & zoeken (search for work)
werk & zwaar (heavy work)

Ontologize Cornetto

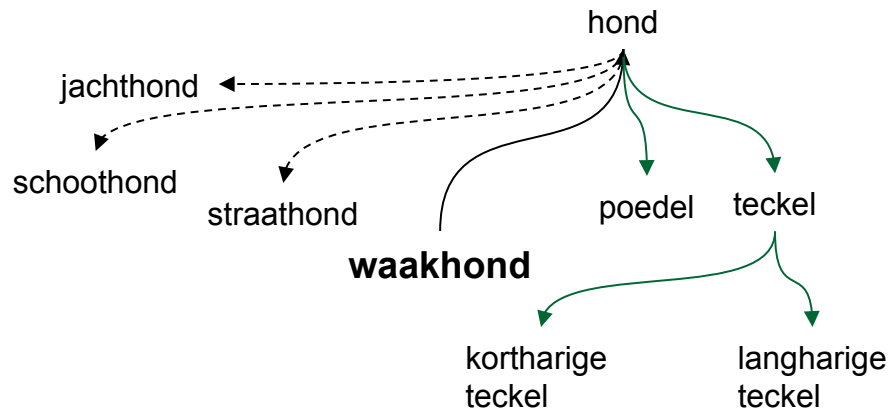
- Identity criteria OntoClean (Guarino & Welty 2002), :
 - **rigidity**: to what extent are properties true for entities in all worlds? You are always a *human*, but you can be a *student* for a short while..
- Hyponyms of *hond* (dog) in DWN:
 - bokser; corgi; loboor; mopshond; pekinees; pointer; spaniël;
 - pup; reu (male dog); teef (female dog)
 - bastaard (bastard dog); straathond (street dog); blindengeleidehond (dog for blind people); bullebijter (tough dog); diensthond (service dog); gashond (for spotting gas); jachthond (hunting dog); lawinehond (finding people after aveline); schoothondje (lap dog); waakhond (watch dog)

Identity criteria applied to DWN

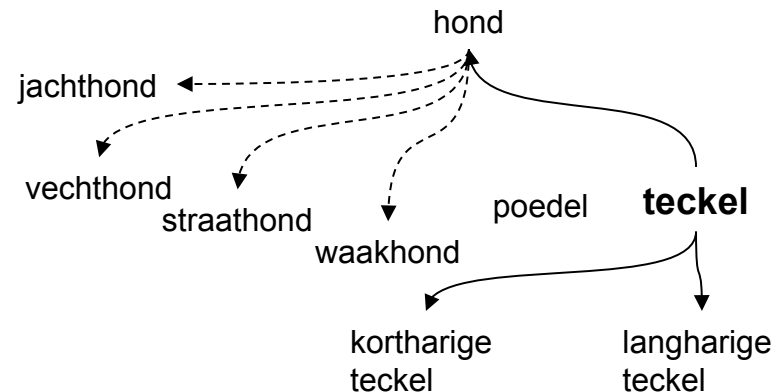
- (Semi-)rigid type hierarchy in the ontology:
 - Canine => PoodleDog; NewfoundlandDog; DalmatianDog, etc.
- Wordnet consists of names for (semi-)rigid dog-types and other words for dogs with roles:
 - poedel = PoodleDog
 - jachthond (?CAN)
 - ⇔ (and
(instance ?CAN Canine)
(instance ?EV Hunting)
(agent ?CAN ?EV)))
- Type hierarchy remains compact and pure

Expansion with pure hyponymy relations

Expansie van rol naar types



Expansie van types naar rollen



antecedent

```
(=>
(instance ?OBJ CorpuscularObject)
(exists (?SUBSTANCE1 ?SUBSTANCE2)
  (and
    (subclass ?SUBSTANCE1 Substance)
    (subclass ?SUBSTANCE2 Substance)
    (material ?SUBSTANCE1 ?OBJ)
    (material ?SUBSTANCE2 ?OBJ)
    (not
      (equal ?SUBSTANCE1 ?SUBSTANCE2))))))
```

Merge.kif
1015-1023

- if ?OBJ is an [instance](#) of corpuscular object,
- then there exist ?SUBSTANCE1 and ?SUBSTANCE2 so that ?SUBSTANCE1 is a [subclass](#) of substance and ?SUBSTANCE2 is a [subclass](#) of substance and ?SUBSTANCE1 is [made](#) of ?OBJ and ?SUBSTANCE2 is [made](#) of ?OBJ and ?SUBSTANCE1 is not [equal](#) to ?SUBSTANCE2

consequent

```
(=>
(instance ?ATTACH Attaching)
(exists (?OBJ)
  (and
    (instance ?OBJ CorpuscularObject)
    (patient ?ATTACH ?OBJ))))

(=>
(instance ?DETACH Detaching)
(exists (?OBJ)
  (and
    (instance ?OBJ CorpuscularObject)
    (patient ?DETACH ?OBJ))))
```

Merge.kif
9266-9271

- if ?ATTACH is an [instance](#) of attaching,
- then there exists ?OBJ so that ?OBJ is an [instance](#) of corpuscular object and ?OBJ is a [patient](#) of ?ATTACH

Merge.kif
9289-9294

- if ?DETACH is an [instance](#) of detaching,
- then there exists ?OBJ so that ?OBJ is an [instance](#) of corpuscular object and ?OBJ is a [patient](#) of ?DETACH

```

(=>
  (and
    (instance ?CUP Cup)
    (contains ?CUP ?STUFF))
  (instance ?STUFF Beverage))

(=>
  (and
    (instance ?DRINK Drinking)
    (patient ?DRINK ?BEV))
  (instance ?BEV Beverage))

(=>
  (and
    (instance ?ORG Organization)
    (attribute ?ORG BeverageAndTobaccoProductManufacturing))
  (exists (?EV ?MEM)
    (and
      (member ?MEM ?ORG)
      (agent ?MEM ?EV)
      (exists (?THING ?MEM2 ?EV2)
        (and
          (instance ?EV Manufacture)
          (or
            (instance ?THING Beverage)
            (instance ?THING TobaccoProduct))
          (patient ?EV ?THING)
          (member ?MEM2 ?ORG)
          (instance ?EV2 Selling)
          (agent ?EV2 ?MEM2)
          (patient ?EV2 ?THING)))))))

```

Mid-level-ontology.kif
1001-1005

- if ?CUP is an [instance](#) of Cup and ?CUP [contains](#) ?STUFF,
- then ?STUFF is an [instance](#) of beverage

Merge.kif
11297-11301

- if ?DRINK is an [instance](#) of drinking and ?BEV is a [patient](#) of ?DRINK,
- then ?BEV is an [instance](#) of beverage

naics.kif 1836-1854

- if ?ORG is an [instance](#) of organization and BeverageAndTobaccoProductManufacturing is an [attribute](#) of ?ORG,
- then there exist ?EV and ?MEM so that ?MEM is a [member](#) of ?ORG and ?EV is an [agent](#) of ?MEM and there exist ?THING and ?MEM2 and ?EV2 so that ?EV is an [instance](#) of manufacture and ?THING is an [instance](#) of beverage or ?THING is an [instance](#) of TobaccoProduct and ?THING is a [patient](#) of ?EV and ?MEM2 is a [member](#) of ?ORG and ?EV2 is an [instance](#) of selling and ?MEM2 is an [agent](#) of ?EV2 and ?THING is a [patient](#) of ?EV2

Language technology opportunities

Text analysis

- Disambiguate word meanings in text => expand training data:
 - horizontal relations
 - frame structure with combinatoric constraints
 - examples
- More refined query expansion to variants and inclusions:
 - Differentiated hyponymy-structures; types versus roles;
- Better cohesion due to lexical inclusion:
 - knippen (to cut) -> schaar (scissors)
 - genezen (to cure) -> ziekte, verwonding (disease, disorder)
- Inferencing:
 - time-dependencies
 - geboorte (get born) -> huwelijk (to marry) -> scheiding (to divorce)-> sterven (to die)
 - visual clues
 - oprichten (to found)-> muziekoptreden (to perform)-> uit elkaar gaan (to split)

Text generation

- Selection of common combinations from alternatives generated by wordnet or ontology:
 - drank (drinks) -> prepareren; maken; brouwen
 - limonade maken; koffie zetten; drankje brouwen
 - smaakvolle soep; pittig eten; sterke koffie;
- Realization of conceptual relations in syntactically correct constructions:
 - behandelen aan verwondingen
 - behandelen voor een ziekte

User scenarios

Applications in Cornetto

- Extension of the database (words & ontology) to the legal domain;
- Improve performance of existing systems:
 - small size database => low recall
 - specialized types of information
 - visual clues,
 - intentions
 - object properties
 - cross-lingual retrieval and classification
 - paraphrasing in summarization and machine translation

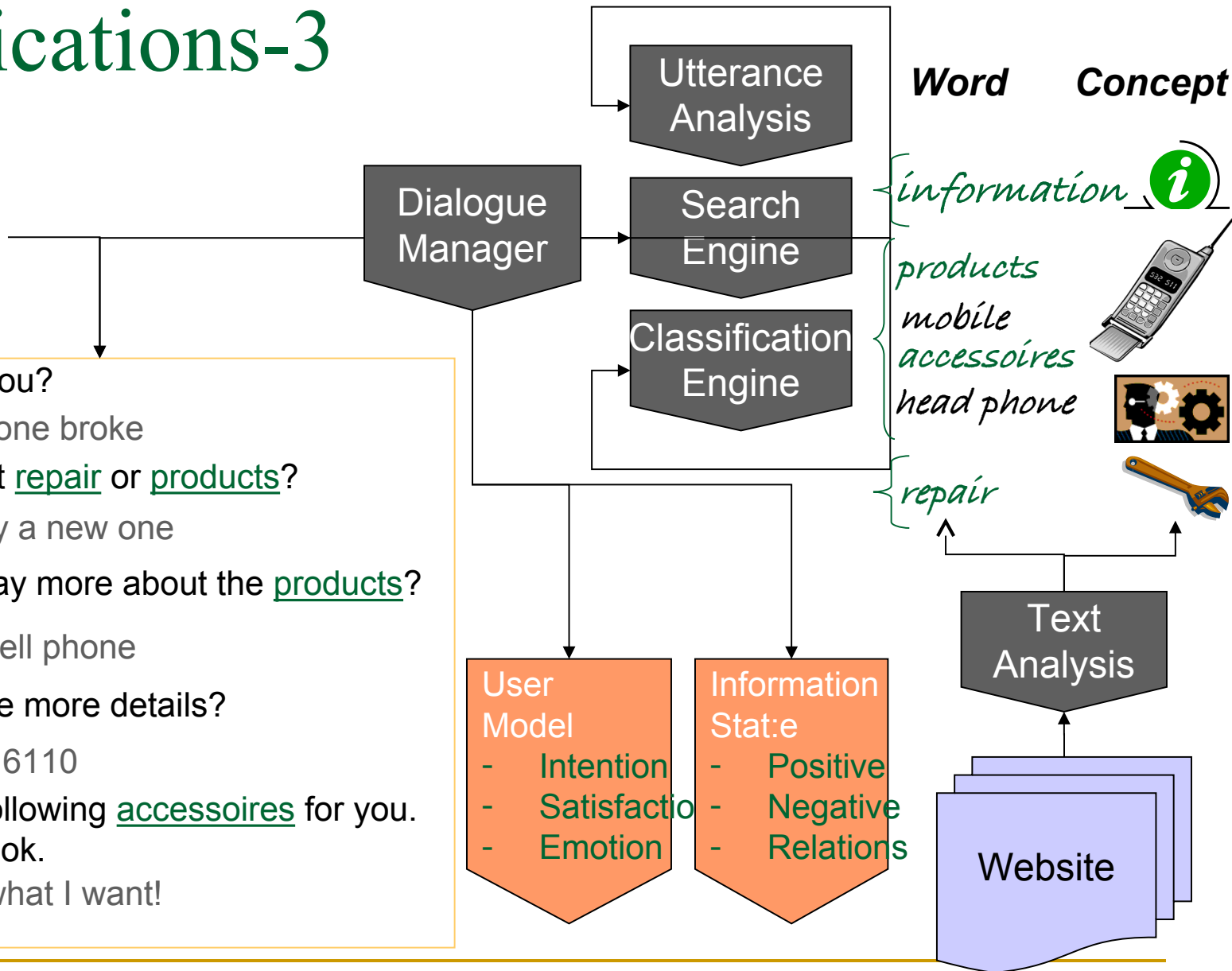
Applications-1

- Summarization:
 - Recognition of duplication
 - Sentence-level paraphrasing
- Entity recognition in visual documents
 - CLASS project, Leuven
 - Image captions
 - Named entity recognition
 - Visual clues:
 - defined for certain types of objects (dog not thoughts)
 - spread to synsets through hypernym chains

Applications-2

- Retrieval in small databases, ISLA, Amsterdam
 - Verkiezingswijzer: 4618 pseudo documents
 - Election themes => hierarchical term lists extracted from party leaflets, e.g.:
 - transport, road pricing, mobility, public, traffic jam, commuting, tunnel, etc..
 - Query expansion and passage retrieval
- Cross-lingual classification, IRION, Delft:
 - IPTC = international press thesaurus committee:
 - IPTC classifier for Dutch, about 1500 classes, 10-20 training documents per class
 - Translate index to other languages (English, German, French, etc.) => bags of words
 - Homogeneous sets of training documents (e.g. weather, football) allow for effective WSD and translation

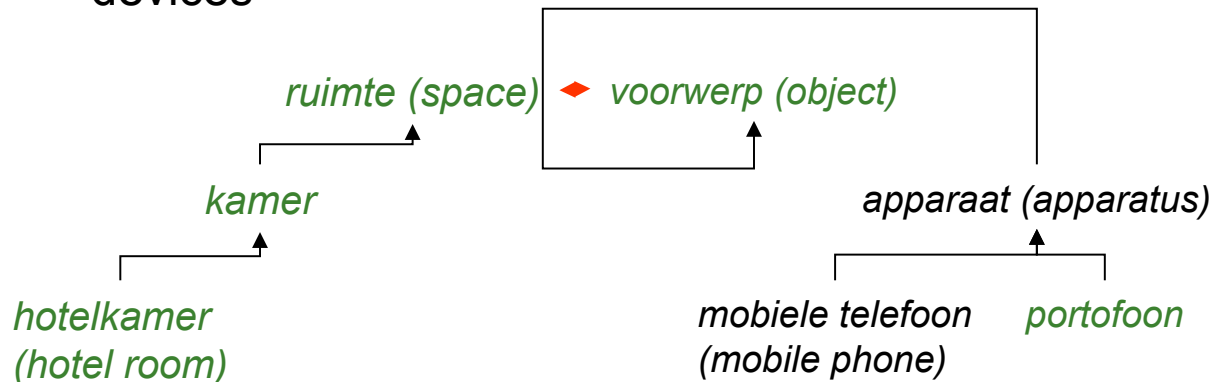
Applications-3



Applications-3: robust dialogue systems

■ Prevents deadlocks:

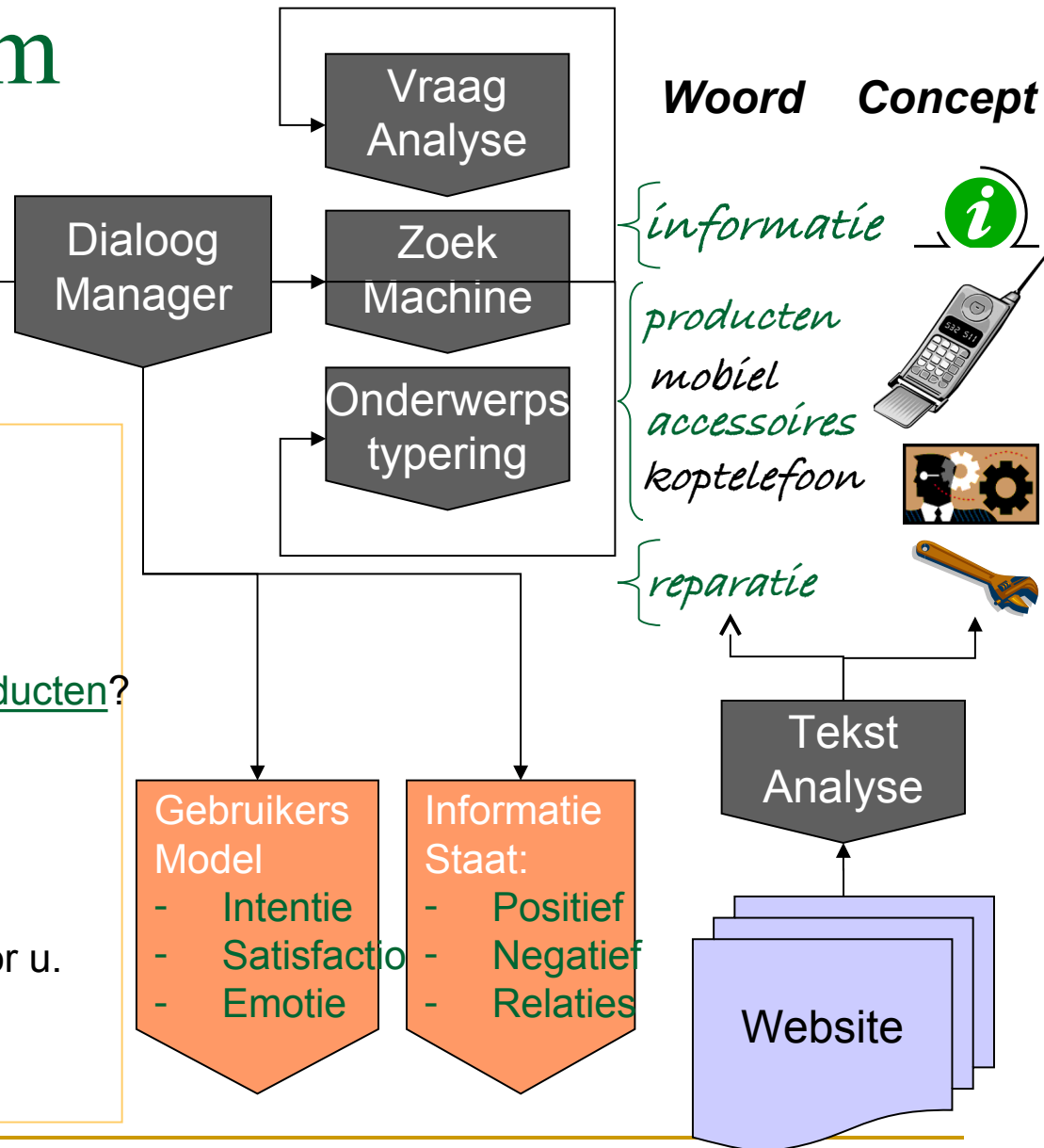
- Detects vagueness and ambiguity (what meaning of *band*?)
- Detects topic changes
- Uses negative information: “*Geen muziekband, ik zoek fietsbanden!*”
- Handles out-of-domain questions in a better way:
 - “We do not have hotel rooms but we do have electronic equipment”
 - “We do not have portophones but we do have other electronic devices”





The end....

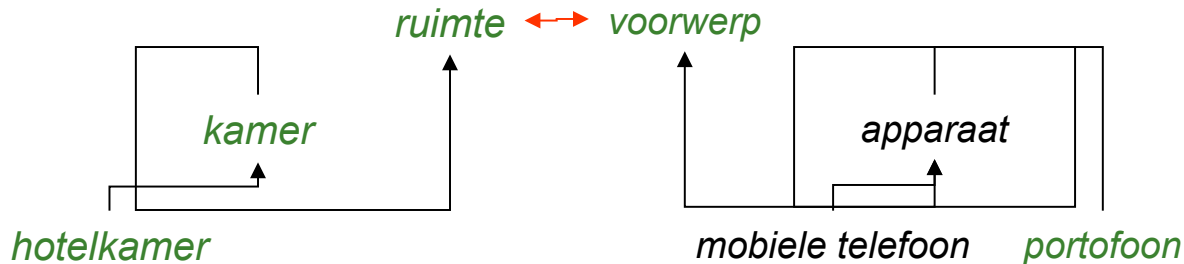
- Bekijk ze eens.



Communicatief dialoogsysteem

- Voorkomt deadlocks:

- ❑ Detecteert vaagheid en ambiguïteit (welke betekenis van *band*?)
- ❑ Detecteert veranderingen van onderwerp
- ❑ Gebruikt negatieve informatie: “*Geen muziekbond, ik zoek fietsbanden!*”
- ❑ Kan out-of-domain vragen aan:
 - “*We hebben geen hotelkamers maar wel elektronische apparaten*”.
 - “*Nee, we hebben geen portofoons maar wel andere elektronische apparaten zoals mobiele telefoons*”



Text retrieval

- Evaluatie raamwerk van EU-projekt MEANING (IST-2001-34460)
- Queries met ambiguïteit en parafraseringen worden gextraheerd, e.g. "police cell" (jail), "cell phone" (mobile), "nerve cell" (neuron).
- Test op verschillende indexen:
 - Index van letterlijke tekst
 - Index van tekst geëxpandeerd met wordnet
 - Index van tekst geëxpandeerd met wordnet na bepalen van betekenis
- Betekenis wordt bepaald door woorden aan domeinen te koppelen en domeinen aan wordnet (Magnini et al 2000): hoge precisie en lage recall;
- Cornetto wordt gebruikt om de recall te vergroten: frame-achtige relaties als er geen domein van toepassing is;

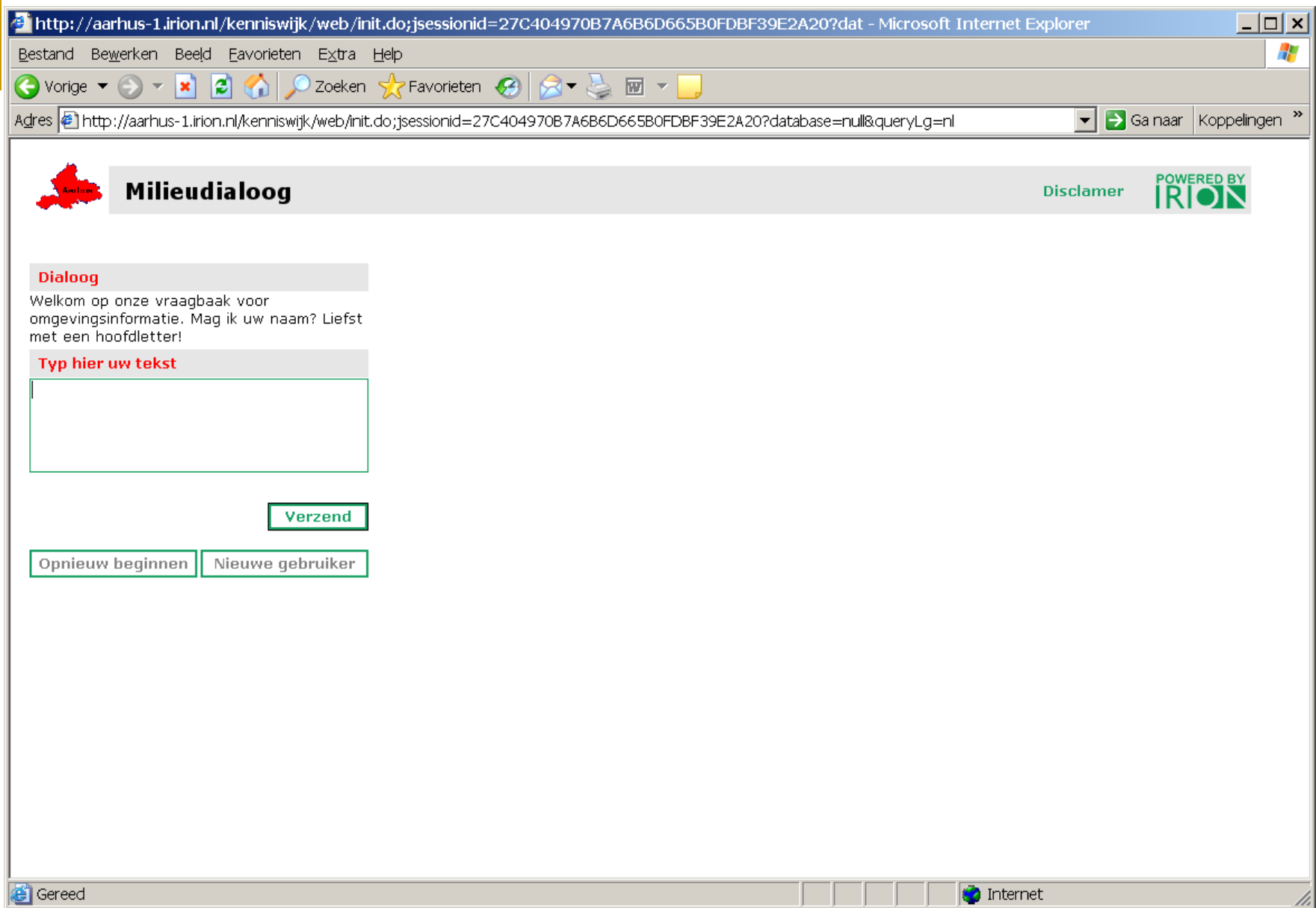
Future scenario's

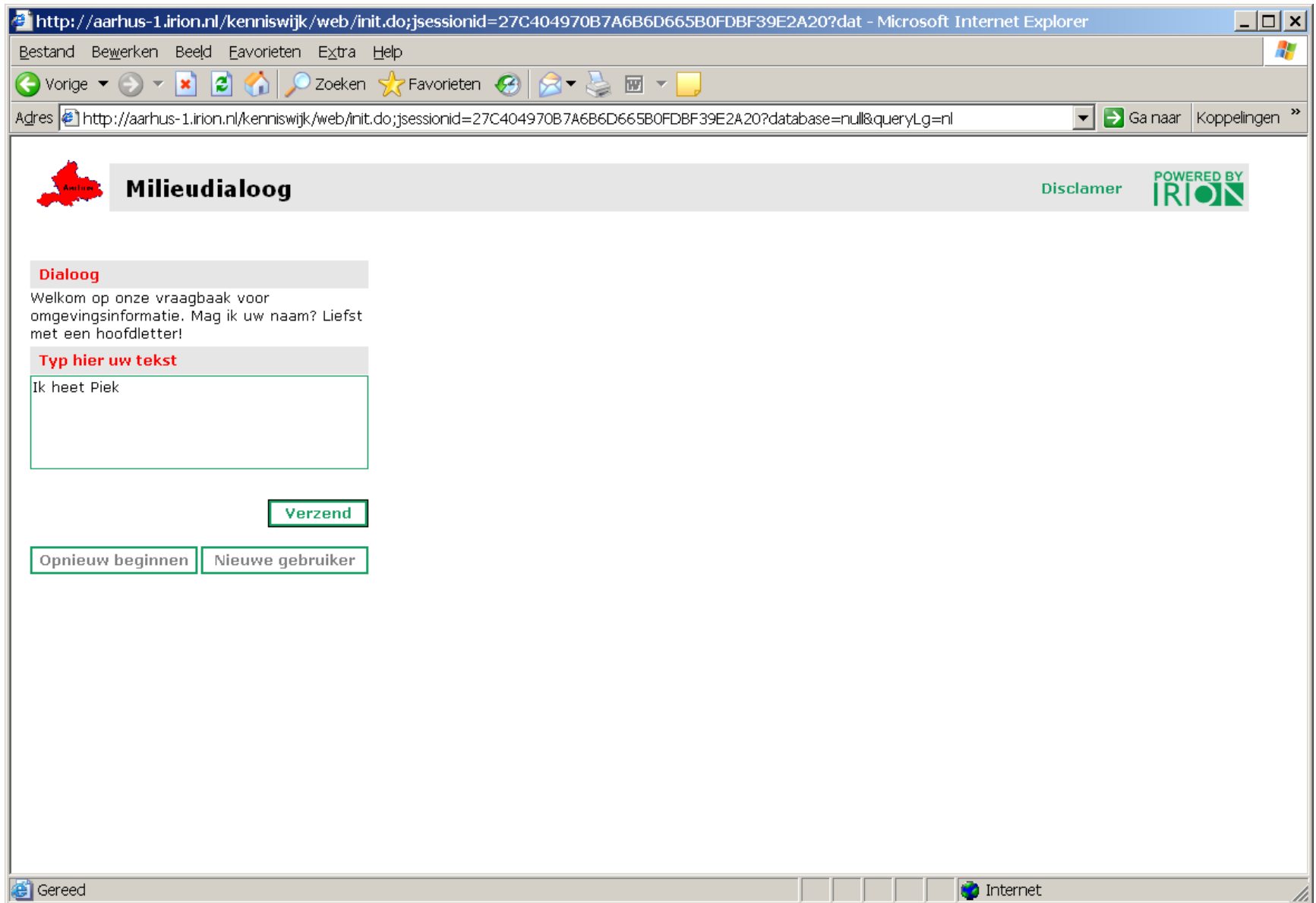
New applications

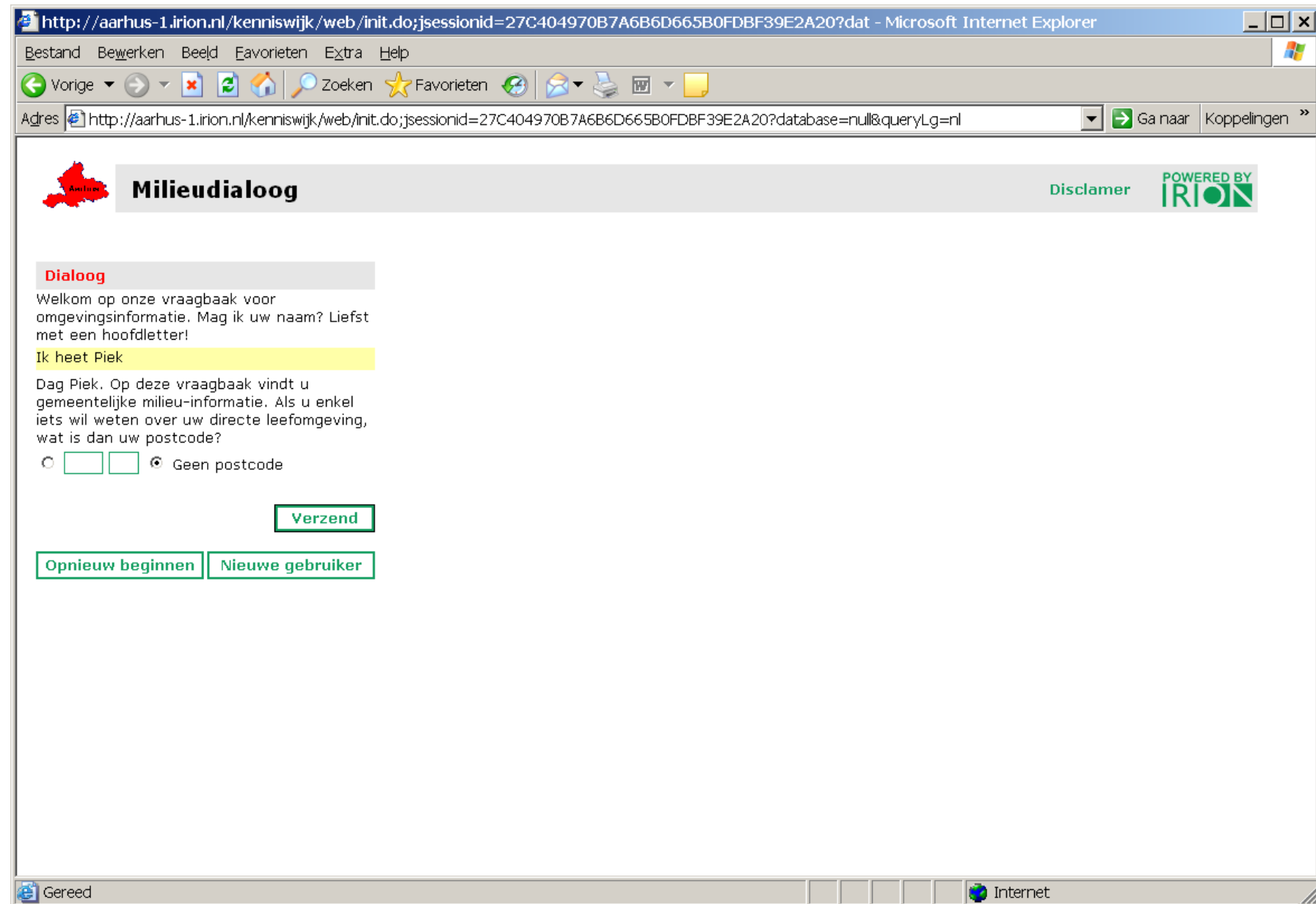
- Data mining
 - Automatic derivation of ontologies
 - Controlled inferencing
- Co-operative dialogue systems:
 - Relation between data objects are important
 - Vagueness and ambiguity are handled
- Summarizers with better cohesion
- Text generation

Beperkingen van gestructureerde data

- Taalgebruik van grote publiek past niet op het taalgebruik van de informatie:
 - Regelgeving:
 - **Voertuigen** op de **openbare weg** moeten worden voorzien van een geldig **kentekenbewijs** en duidelijke **markeringen**;
 - Gebruikersvraag:
 - Ik heb een **old-timer** die ik haast nooit gebruik. Moet die ook een **kentekenplaat** en **lichten** hebben als ik die op **straat** parkeer?
- Eindgebruikers willen niet een complexe indeling volgen/leren/kennen maar hun eigen perspectief









http://aarhus-1.iron.nl/kenniswijk/web/init.do;jsessionid=27C404970B7A6B6D665B0FDBF39E2A20?dat - Microsoft Internet Explorer

Bestand Bewerken Beeld Favorieten Extra Help

Vorige Zoeken Favorieten

Adres http://aarhus-1.iron.nl/kenniswijk/web/init.do;jsessionid=27C404970B7A6B6D665B0FDBF39E2A20?database=null&queryLg=nl Ga naar Koppelingen »

 **Milieudialoog** Disclamer **POWERED BY IRION**

Dialoog

Welkom op onze vraagbaak voor omgevingsinformatie. Mag ik uw naam? Liefst met een hoofdletter!

Ik heet Piek

Dag Piek. Op deze vraagbaak vindt u gemeentelijke milieu-informatie. Als u enkel iets wil weten over uw directe leefomgeving, wat is dan uw postcode?

Geen postcode.

Dank u, ik zal over de hele gemeente gaan zoeken. Ik werk als volgt: u stelt een vraag, ik leid u naar een rubriek en daarbinnen naar een document waarin u het antwoord kunt lezen. Wat wilt u weten?

Typ hier uw tekst

Kan fijn stof tot kanker leiden?

Verzend

Opnieuw beginnen **Nieuwe gebruiker**

Gereed Internet

http://aarhus-1.irion.nl/kenniswijk/web/newSession.do - Microsoft Internet Explorer

Bestand Bewerken Beeld Favorieten Extra Help

Vorige Zoeken Favorieten

Adres http://aarhus-1.irion.nl/kenniswijk/web/newSession.do Ga naar Koppelingen »



Milieudialoog

[Disclaimer](#) **POWERED BY IRION**

Dialog

Leuk dat u onze vraagbaak voor omgevingsinformatie bezoekt. Wat is uw naam? Graag met een hoofdletter!

Piek

Goedendag Piek. Deze vraagbaak bevat lokale milieu-informatie. Als u enkel iets wilt vragen over uw directe leefomgeving, wat is dan uw postcode?

Geen postcode.

Oké, ik ga voor u op zoek in de hele gemeente. Ik werk als volgt: u stelt een vraag, ik leid u naar een rubriek en daarbinnen naar een document waarin u het antwoord kunt lezen. Wat wilt u vragen?

Kan fijn stof tot kanker leiden?

Ik kan helaas zelf geen vragen beantwoorden Piek. Wel kan ik wat informatie voor u opzoeken.

Wilt u iets weten over rubriek **Lucht** of **Gezondheid**?

Typ hier uw tekst

Verzend

Opnieuw beginnen **Nieuwe gebruiker**

Internet

http://aarhus-1.irion.nl/kenniswijk/web/newSession.do - Microsoft Internet Explorer

Bestand Bewerken Beeld Favorieten Extra Help

Vorige Zoeken Favorieten

Adres http://aarhus-1.irion.nl/kenniswijk/web/newSession.do Ga naar Koppelingen »

 **Milieudialoog** Disclaimer POWERED BY IRION

Dialoog

Leuk dat u onze vraagbaak voor omgevingsinformatie bezoekt. Wat is uw naam? Graag met een hoofdletter!

Piek

Goedendag Piek. Deze vraagbaak bevat lokale milieu-informatie. Als u enkel iets wilt vragen over uw directe leefomgeving, wat is dan uw postcode?

Geen postcode.

Oké, ik ga voor u op zoek in de hele gemeente. Ik werk als volgt: u stelt een vraag, ik leid u naar een rubriek en daarbinnen naar een document waarin u het antwoord kunt lezen. Wat wilt u vragen?

Kan fijn stof tot kanker leiden?

Ik kan helaas zelf geen vragen beantwoorden Piek. Wel kan ik wat informatie voor u opzoeken.

Wilt u iets weten over rubriek **Lucht** of **Gezondheid**?

Ga verder in Gezondheid

Oké Piek, ik zal voor u in rubriek **Gezondheid** kijken...

Ik twijfel sterk, maar misschien heeft u toch iets aan rubriek **Lucht en gezondheid**. Klik erop.

Typ hier uw tekst

Verzend

Opnieuw beginnen **Nieuwe gebruiker**

Internet

http://aarhus-1.irion.nl/kenniswijk/web/newSession.do - Microsoft Internet Explorer

Bestand Bewerken Beeld Favorieten Extra Help

Vorige Zoeken Favorieten

Adres http://aarhus-1.irion.nl/kenniswijk/web/newSession.do Ga naar Koppelingen

Milieudialogo

Disclaimer POWERED BY IRION

omgevingsinformatie bezoekt. Wat is uw naam? Graag met een hoofdletter!

Piek

Goedendag Piek. Deze vraagbaak bevat lokale milieu-informatie. Als u enkel iets wilt vragen over uw directe leefomgeving, wat is dan uw postcode?

Geen postcode.

Oké, ik ga voor u op zoek in de hele gemeente. Ik werk als volgt: u stelt een vraag, ik leid u naar een rubriek en daarbinnen naar een document waarin u het antwoord kunt lezen. Wat wilt u vragen?

Kan fijn stof tot kanker leiden?

Ik kan helaas zelf geen vragen beantwoorden Piek. Wel kan ik wat informatie voor u opzoeken.

Wilt u iets weten over rubriek **Lucht** of **Gezondheid**?

Ga verder in **Gezondheid**

Oké Piek, ik zal voor u in rubriek **Gezondheid** kijken...

Ik twijfel sterk, maar misschien heeft u toch iets aan rubriek **Lucht en gezondheid**. Klik erop.

Ga verder in **Lucht en gezondheid**

Gelukkig Piek, ik heb iets voor u. Zie rechts.

Typ hier uw tekst

Verzend

Opnieuw beginnen Nieuwe gebruiker

Gevonden: 8 document(en). Klik op plusje (+) om treffers te bekijken.

Alles tonen Alles verbergen

1. CROW: Wegen naar schonere lucht (48 pag.) Metadata

88% Pag. 13 Rubriek: Lucht en gezondheid; Postcode: 6533D; verbindingen tussen zuurstof en **stikstof**. De voornaamste zijn stikstofmonoxide(NO) en maximaal twee jaar eerder overlijden dan wanneer zij niet waren blootgesteld. **Fijn stof** Kan diep in de longen en verhoogde hart slag. **Mogelijk kankerverwekkend**.- Vervroegde sterfte aan alle doodsoorzaken.- **Pag in HTML**

60% Pag. 14 Rubriek: Lucht en gezondheid; Postcode: 6533D; De bijdrage van de weg aan de concentraties **fijn stof** en stikstofdioxide kan wel leiden tot lokale overschrijdingen. Voor een overzicht van de werking van **fijn stof**, stikstofdioxide en ozon en blootstelling aan een hoge concentratie van **stoffen** [6]. Hierbij is onderscheid gemaakt tussen effecten bij kortdurende **Pag in HTML**

60% Pag. 38 Rubriek: Lucht en gezondheid; Postcode: 6533D; bio-ethanol hebben geen effect op **fijn stof** en stikstofdioxide concentraties. Effecten op een afname van de concentraties stikstofdioxide en **fijn stof** achter het scherm. Het positieve effect neemt toe met de hoogte van het scherm. **Pag in HTML**

60% Pag. 36 Rubriek: Lucht en gezondheid; Postcode: 6533D; Een bus met een aardgasmotor heeft bijvoorbeeld een duidelijk lagere uitstoot **stikstofoxiden** dan een bus met een diesel motor. Ook de uitstoot van **fijn stof** is lager. Het effect op de bijdrage van het verkeer aan **Pag in HTML**

2. Provincie Gelderland: Rapportage luchtkwaliteit 2004 (23 pag.) Metadata

3. Milieu- en Natuurplanbureau: Fijn stof nader bekeken (64 pag.) Metadata

4. Voorontwerp Uitvoeringsprogramma verbetering luchtkwaliteit Nijmegen 2005-2010 (112 pag.) Metadata

5. Rapportage Besluit Luchtkwaliteit 2002 (65 pag.) Metadata

6. Rapportage Besluit Luchtkwaliteit 2001 (65 pag.) Metadata

7. Milieu- en Natuurplanbureau: Luchtkwaliteit en Gezondheid (1 pag.) Metadata

8. Rapportage luchtkwaliteit 2004 (23 pag.) Metadata

Internet



The end....