De automatische analyse van subjectiviteit en causale samenhang in tekst

Wilbert Spooren

Met dank aan:
Erwin Komen, Micha Hulsbosch, Iris Hofstra, Patrick Sonsma
Jet Hoek, Ted Sanders
Clariah

Radboud Universiteit

Subjectiviteit en causale connectieven in tekst

The temperature rose because the sun was shining.
The neighbors must be away, because the lights are out.

Overeenkomsten

- “achterwaartse causaliteit”: Gevolg-Oorzaak
- beide veronderstellen een implicatierelatie $P \rightarrow Q$

Verschil:

rapporteren van externe werkelijkheid (OBJECTIEF)
versus
conclusie van een spreker/schrijver (SUBJECTIEF)

Achterwaartse causaliteit in het Nederlands

De temperatuur steeg doordat de zon scheen.
FYSIEKE GEVOLG-OORZAAK-relatie (non-volitionele inhoudsrelatie)

Jan ging naar huis omdat hij ziek was.
GEVOLG-REDEN-relatie (volitionele inhoudsrelatie)

De buren zijn waarschijnlijk weg want de lichten zijn uit.
BEWERING-ARGUMENT (epistemische relatie)

Moet iemand nog naar de wc want we gaan zo weg.
SPEECH ACT

Vergelijking van causale connectieven in NL in drie genres

- geschreven krantenteksten
  - hoge mate van editing, zorgvuldig gepland
- gesproken conversaties
  - spontaan, face-to-face, lage graad van planning
- chat
  - spontaan, lage graad van planning, weinig feedback van conversationele partner

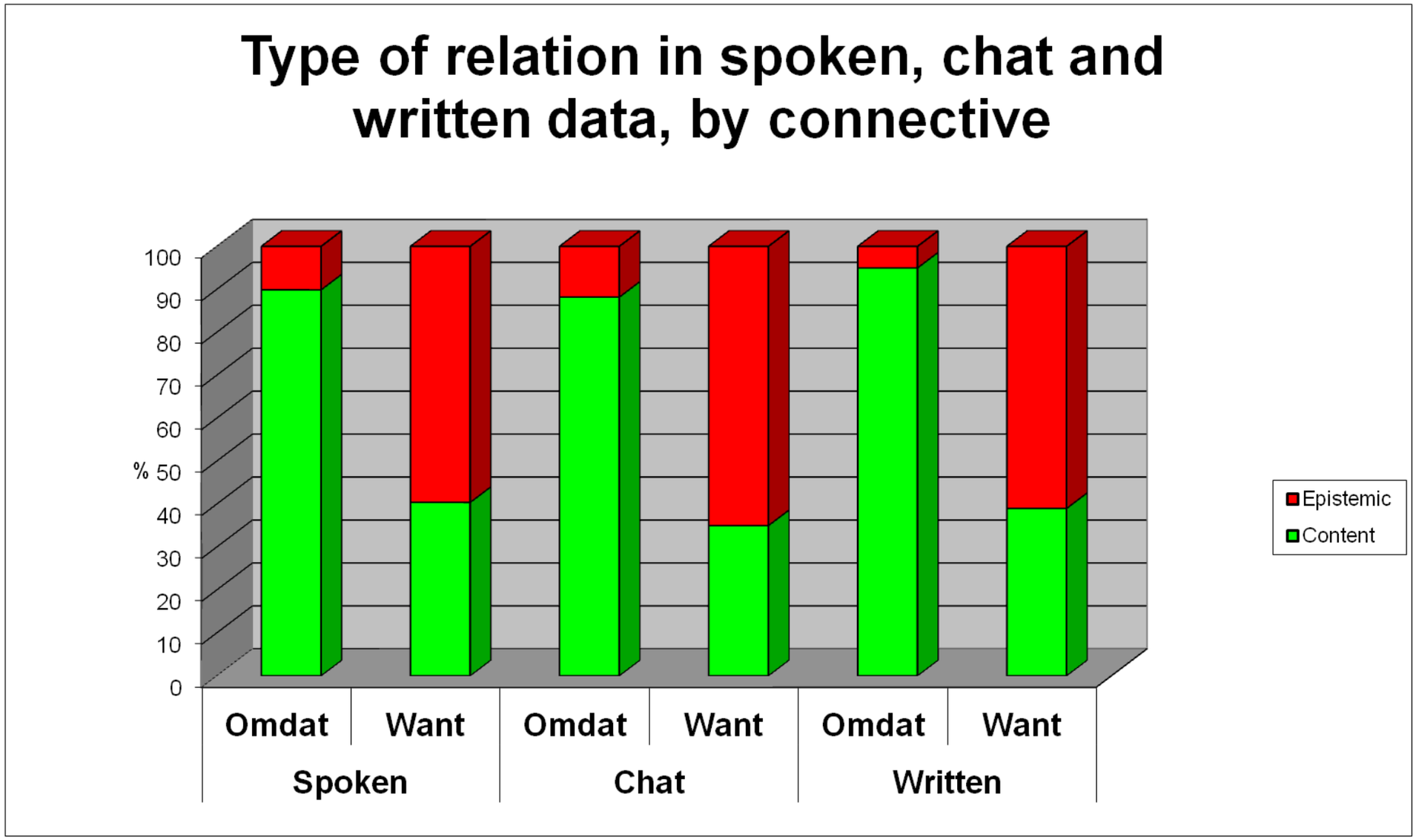
- Onderzoeksvragen:
  - komen want en omdat voor in verschillende omgevingen?
  - is de distributie van want en omdat genre-afhankelijk?

Sanders & Spooren (2015). Causality and subjectivity in discourse: The meaning and use of causal connectives in spontaneous conversation, chat interactions and written text. *Linguistics, 53* (1), 53-92.

Typische onderzoeksopzet

- Handmatige analyse van relatief kleine steekproeven van fragmenten
- 100 vbb van *omdat* en *want* per genre
- Elk fragment geanalyseerd op een groot aantal eigenschappen, waaronder subjectiviteit

Voorbeeld van een resultaat



Voordelen van automatische analyse

- Minder afhankelijk van handmatige analyse
  - betrouwbaarheid ↑
  - schaalgrootte ↑
  - aantal genres ↑

ACAD: Automatic Coherence Analysis of Dutch

Universiteit Utrecht

Radboud Universiteit

Doelstellingen van ACAD

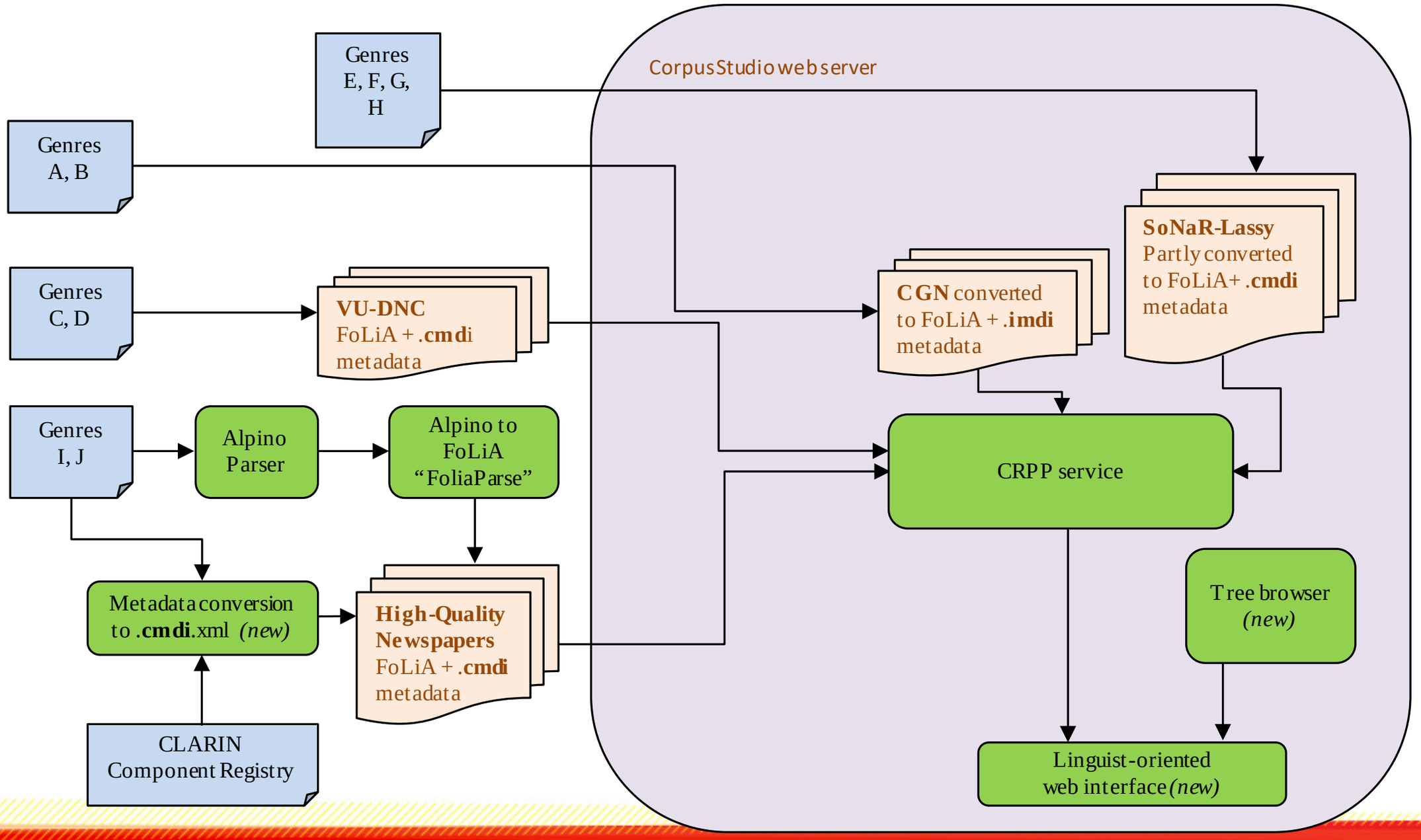
- Zoekinterface, op basis van bestaande Clariah-componenten
  - corpora zoals SoNaR, VU-DNC, CGN
  - parsers zoals Alpino
  - formaten zoals Folia
  - zoekfaciliteiten zoals CorpusStudio
- Geavanceerde zoekopdrachten voor computationeel niet ingewijde tekstwetenschappers
  - achter de schermen vertaald naar XQuery
- Uitbreiden van beschikbare corpora
  - krantendata (NRC en NRC.nl) uit verschillende genres (blogs, hard nieuws, opinie, achtergrond)
  - WhatsAppdata van verschillende leeftijdsgroepen (13/14, 20-25)

ACAD: Automatic Coherence Analysis of Dutch

Universiteit Utrecht

Radboud Universiteit

Opzet ACAD


Genres
E, F, G, H
Genres
A, B
Genres
C, D
Genres
I, J
Alpino
Parser
Alpino to
FoLiA
“FoliaParse”
Metadata conversion
to .cmdi.xml (new)
CLARIN
Component Registry
VU-DNC
FoLiA + .cmdi
metadata
High-Quality
Newspapers
FoLiA + .cmdi
metadata
CorpusStudio web server
CGN converted
to FoLiA + .imdi
metadata
SoNaR-Lassy
Partly converted
to FoLiA+ .cmdi
metadata
CRPP service
Tree browser
(new)
Linguist-oriented
web interface (new)

ACAD: Automatic Coherence Analysis of Dutch

Hoe werkt het?

1. *Voorbereiden van deelcorpora*.
   - verzamelen en converteren van bestaande subcorpora
   - toevoegen van nieuwe corpora

2. *Identificatie van relevante gevallen*.
   - formuleren van zoekopdrachten om alle causale connectieven in de subcorpora te vinden.
   - uitdaging: onderscheid ‘false positives’ (bv. voorzetsel *om*) van ‘true positives’.

Hoe werkt het?

3. *Identificatie van S1 en S2*.
   - omvang van de tekstsegmenten verbonden door causale connectieven moet automatisch vastgesteld worden.
   - uitdaging: incomplete S1/S2, ingebedde constructies, constructies verspreid over verschillende zinnen en/of sprekers (CGN)

4. *Identificatie van de richting van de causale verbinding*.
   - voorwaarts, achterwaarts?
   - hangt af van (i) type connectief (onderschikkend/nevenschikkend voegwoord, bijwoord), (ii) semantiek van het connectief en (iii) grammaticale positie (“conn-S1-S2” versus “S1-conn-S2”)

Hoe werkt het?

5. *Vaststellen van de subjectiviteit van de verbinding*.
   - hoe subjectief is S1/S2? thematische tekstanalyes (Bestgen et al., 2006).
   - de zoekinterface bepaalt kenmerken van de verbinding zoals het aantal subjectieve bijvoeglijk naamwoorden en bijwoorden in S1/S2.

ACAD: Automatic Coherence Analysis of Dutch

De zoekinterface

Research project: **Want_2**

Purpose: Find "want" in any position. Identify S1 and S2. Count number of subjective words in S1.
Main search: Word(s)
Created: 2 september 2017 16:30
Saved: 9 januari 2018 18:05
Shared with: admin_user (reading) seeker_user (reading)

Edit | Summary | Search...

About
Information on this project

Define
Indicate what elements we are interested in

Fine-tune
Specify detailed requirements of a hit

Add features
Specify the features to be added to a search hit

ACAD: Automatic Coherence Analysis of Dutch

Verfijning van de zoekopdracht

Fine-tune

Specify detailed requirements of a hit

<< Overview

Show summary

The search program will attempt to find the words or constituents specified in `Search` (see `Overview`).

But if there are additional `conditions` before the results are of interest, these conditions need to be specified.

The `conditions` may make use of:

- Fixed `global` variables.
  A global variable is a piece of text. Whenever a particular word or clause is used several times in the `data-dependant` variables or in the `conditions`, think of replacing that text by defining a global variable.
- Variables that are `data-dependant`.
  The value of data-dependant variables is defined separately for each word or constituent that is being searched. Data-dependent variables can be of any kind, e.g.: boolean, string, constituent.

Fixed

Global variables

Variables

Data-dependant

Conditions

When is it a hit?

ACAD: Automatic Coherence Analysis of Dutch

Wat komt er in de output

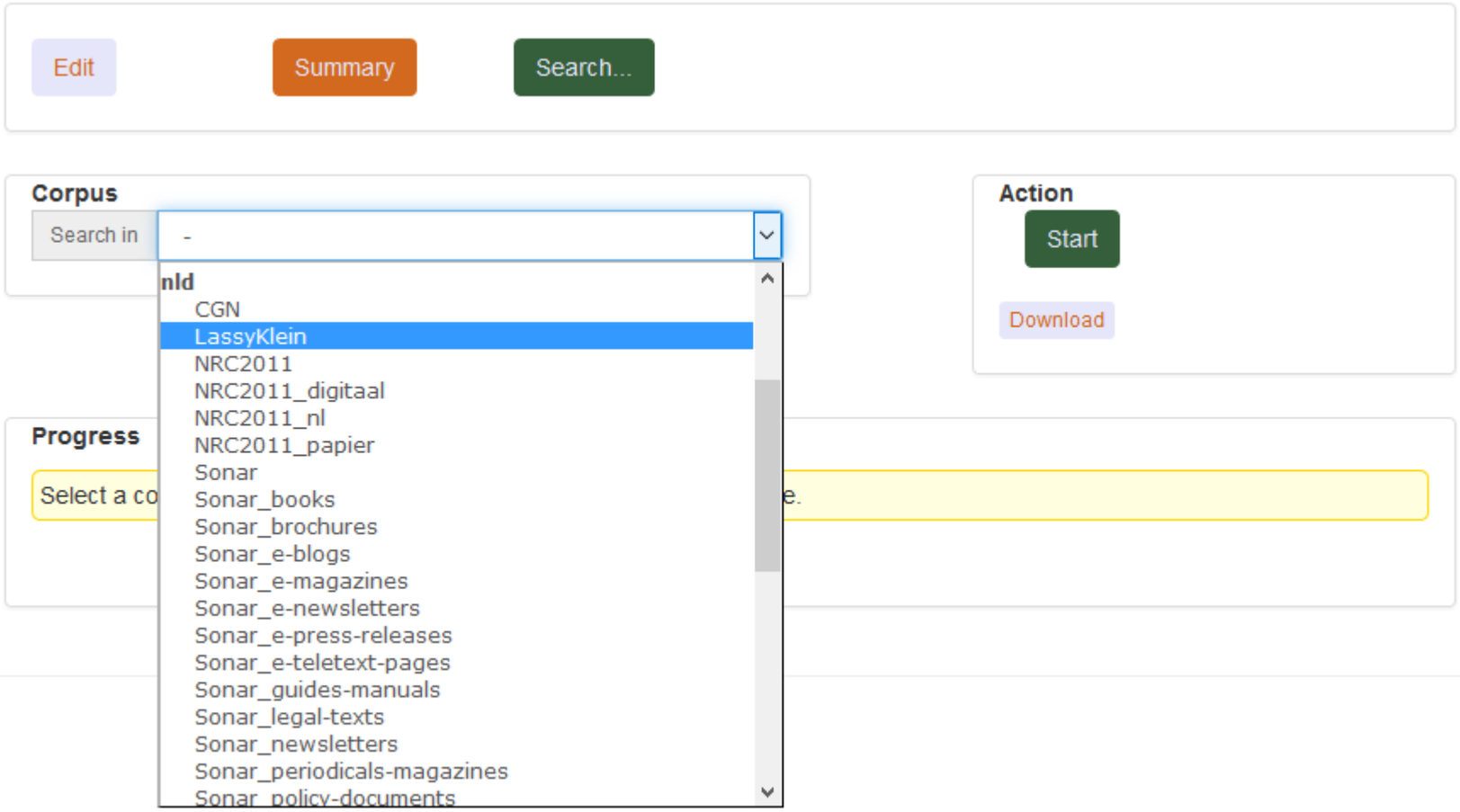
Add features

Specify the features to be added to a search hit

11	s1	The words in S1	Function	get_text (1 args) edit summary...	yes	☐
12	s1_last	The last word in S1	Function	get_text (1 args) edit summary...	yes	☐
13	s1_wordcats	List of word categories in S1	Function	get_wordcat_list (1 args) edit summary...	yes	☐
14	s2	The words in S2	Function	get_text (1 args) edit summary...	yes	☐
15	s1nodeverbal	WHether the S1 contains a verb	Data-dependant variable	s1nodeVerbal	yes	☐
16	s2nodeverbal	Whether the S2 contains a verb	Data-dependant variable	s2nodeVerbal	yes	☐
17	cpModAnc_cat	Category of [cpModAnc] variable	Function	get_cat (1 args) edit summary...	yes	☐
18	lemmas_s1_neutral	All lemma's of neutral adjective in s1	Function	get_text (1 args) edit summary...	yes	☐
19	count_s1_neutral	Number of neutral adjectives in s1	Function	wrd_count (1 args) edit summary...	yes	☐
20	lemmas_s1_subjective	All lemma's of subjective adjective/adverb in s1	Function	get_text (1 args) edit summary...	yes	☐
21	count_s1_subjective	Number of subjective adjectives/adverbs in s1	Function	wrd_count (1 args) edit summary...	yes	☐
22	count_s1	Number of words in S1	Function	wrd_count (1 args) edit summary...	yes	☐

ACAD: Automatic Coherence Analysis of Dutch

De zoekopdracht: kies subcorpora



ACAD in actie

Thematische tekstanalyse

- Hypothese
  - S1 in want-verbinding bevat meer subjectieve woorden dan S1 in omdat-verbinding

- Methode
  - gold1000 lexicon van subjectieve adjectieven (De Smedt & Daelemans, 2012) [1044 adjectieven, elk woord beoordeeld door zeven beoordelaars]
  - subjectief: adjectieven met een score van 0.7 of hoger (voor elk van de betekenissen van het adjectief)
  - voor S1 is het aantal subjectieve adjectieven in S1 en S2 geteld voor want en omdat

- Subcorpus: Lassieklein

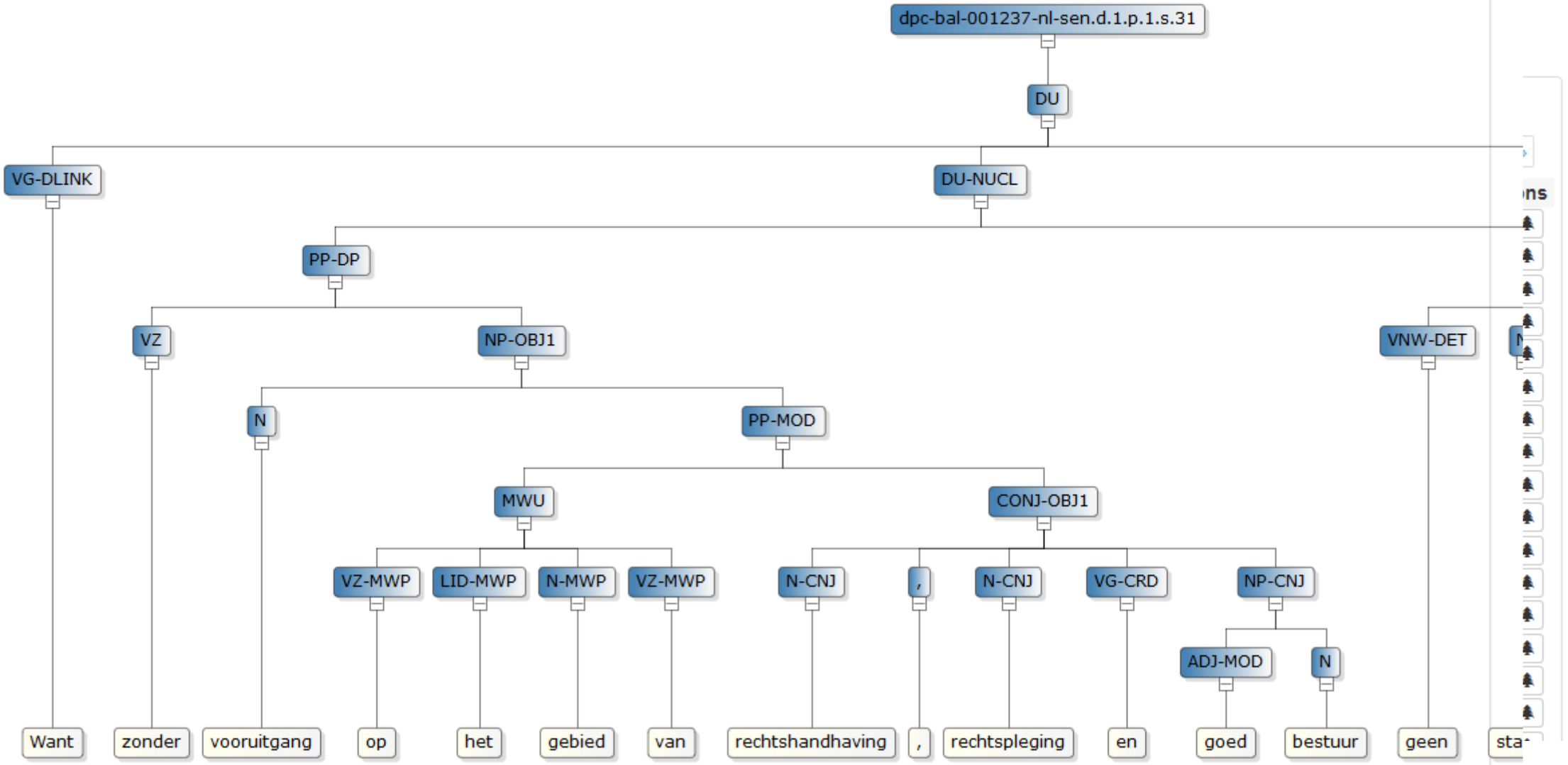
Search results

Search result line: 1 (472)

<< Result center

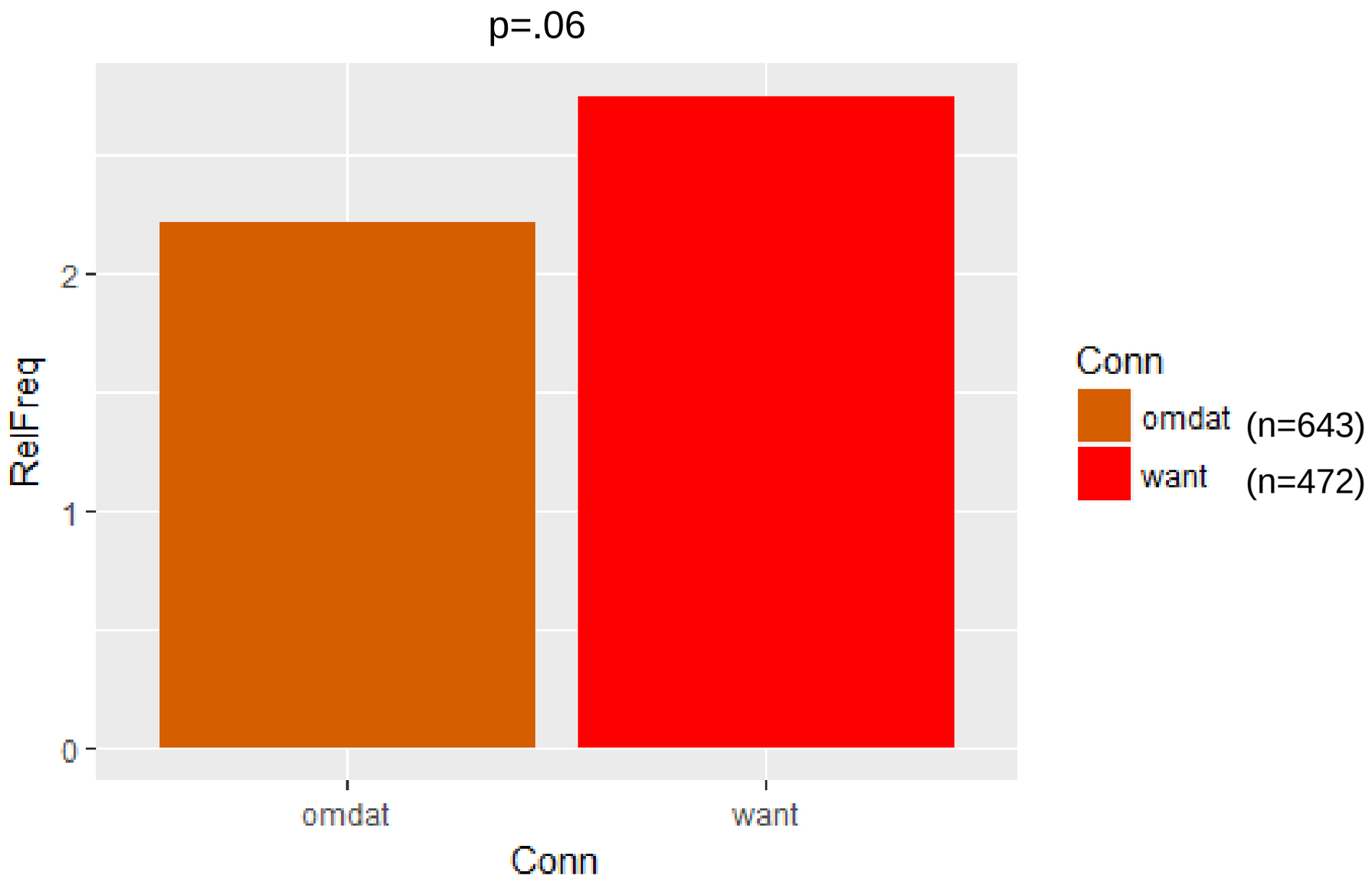
<< Back

Informatie...



ACAD: Automatic Coherence Analysis of Dutch

Resultaten



Volgende stappen

- precision en recall optimaliseren
- meer vbb van zoekvragen
- corpora toevoegen/controleren
- handleidingen maken