The automatic analysis of subjectivity and causal coherence in text

Wilbert Spooren
Ted Sanders

Thanks to
Erwin Komen, Micha Hulsbosch, Henk van den Heuvel, Iris Hofstra, Patrick Sonsma
Jet Hoek
Clariah

Universiteit Utrecht

Radboud Universiteit

Subjectivity and causal connectives in text

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

Similarities

- “backward causality”: Consequence-Cause
- both imply an implicational relation $P \rightarrow Q$

Difference:

report of external reality (OBJECTIVE)
versus
conclusion of a speaker/author (SUBJECTIVE)

Topic 4: Backward causality in Dutch

The temperature rose doordat the sun was shining.
CONSEQUENCE-CAUSE, non-volitional content

Jan went home omdat he was ill.
CONSEQUENCE-CAUSE volitional actions

The neighbors must be away want the lights are out.
CLAIM-ARGUMENT / EPISTEMIC

Does anybody need to go to the bathroom? want we are leaving!
SPEECH ACT

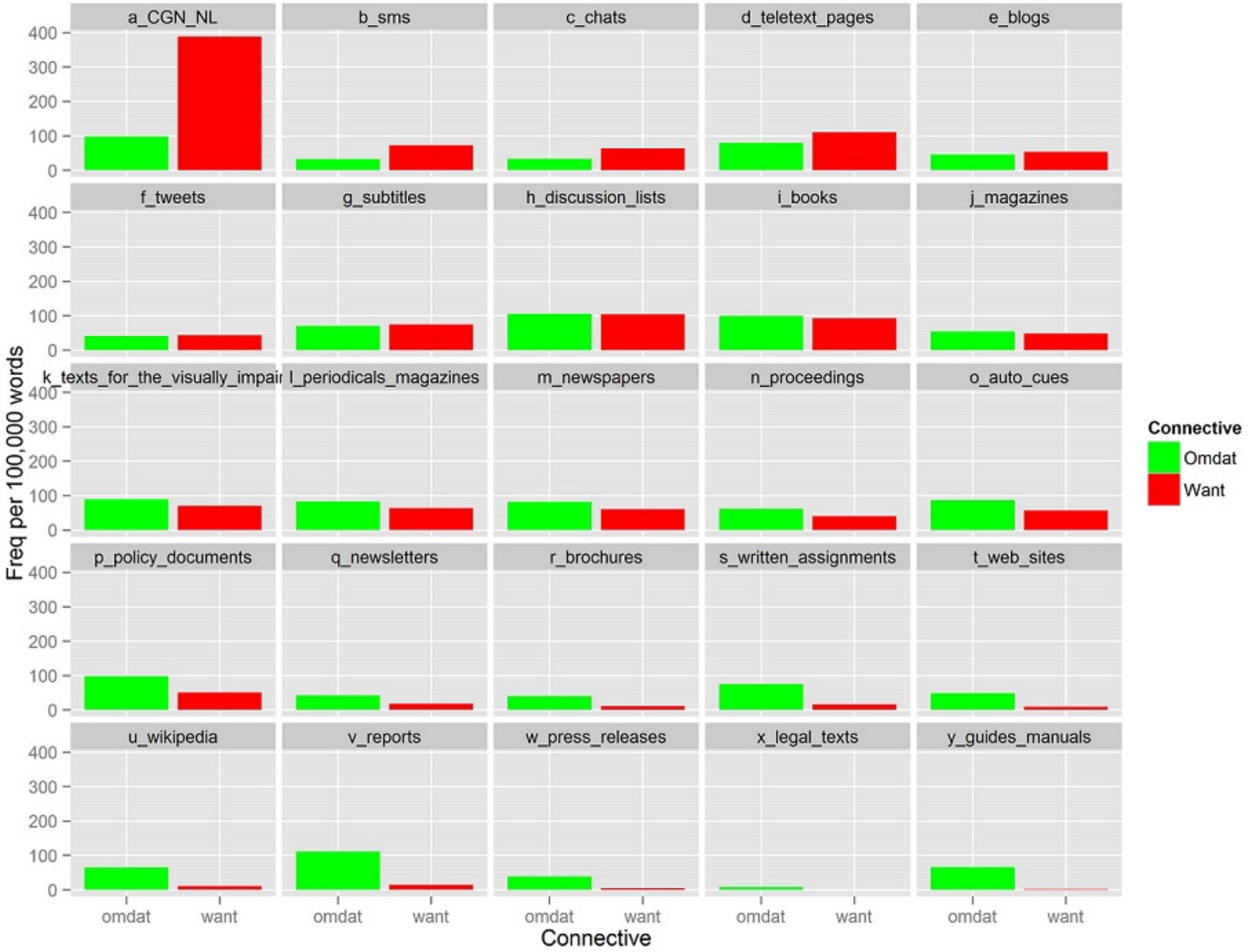
Comparion of causal connectives in Dutch in three genres

- written newspaper texts
  - high degree of editing, carefully planned
- spoken conversations
  - spontaneous, face-to-face, low degree of planning
- chat
  - spontaneous, low degree of planning, less feedback from conversational partner (no paralinguistic cues)

- Research questions:
  - do want and omdat occur in different types of context?
  - is the distribution of want and omdat genre-dependent?

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.

Possible role of genre



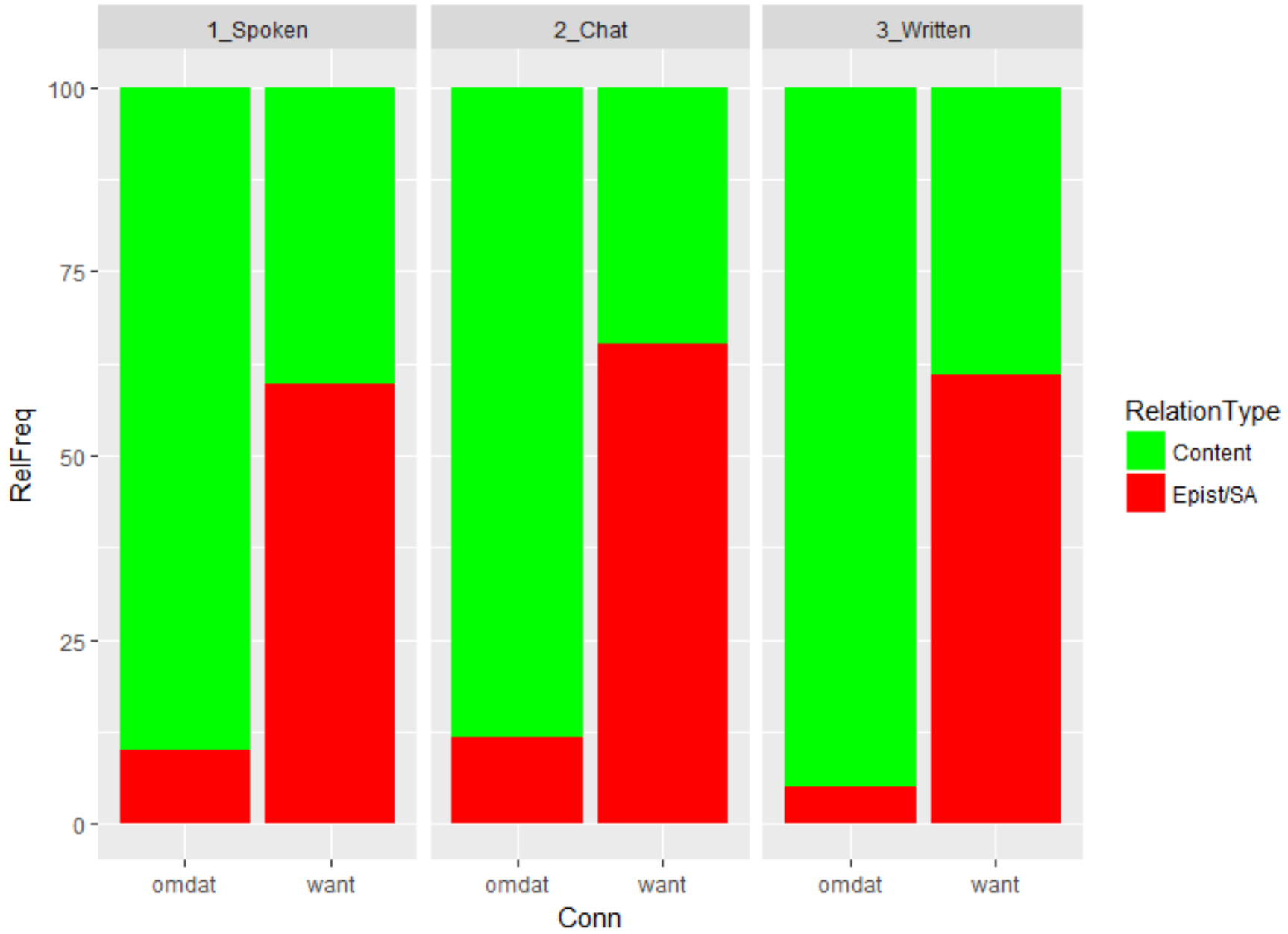
Thanks to
Tijn Schmitz

Typical research design

- Manual analysis of relatively small samples of fragments
- 100 instances of *omdat* and *want* per genre
- Each fragment is analysed on a large number of properties, among which subjectivity

Example result

Relation type as function of connective and genre



Using automatic analyses

- Less dependent on manual analyses
  - higher reliability
  - larger samples
  - larger number of genres

ACAD: Automatic Coherence Analysis of Dutch

Universiteit Utrecht

Radboud Universiteit

Universiteit Utrecht

Radboud Universiteit

Goals of ACAD

ACAD: Automatic Coherence Analysis of Dutch

Universiteit Utrecht Radboud Universiteit

- Build a search interface, on the basis of existing Clariah components
  - corpora like SoNaR, VU-DNC, CGN
  - parsers like Alpino
  - formats like Folia
  - search facilities like CorpusStudio
- Make it possible to formulate sophisticated search queries for computationally uninitiated discourse analysts
  - translated into XQuery in the backend
- Make analyses reproduceable (and consequently more transparent)
- Extend the available corpora
  - newspaper texts (NRC and NRC.nl) from different genres (hard news, opinion, background stories) on related topics
  - WhatsAppdata of different age groups (13/14, 20-25)

How does it work?

1. *Preparation of corpora*
   - collecting and converting existing corpora
   - adding new corpora (including metadata)

2. *Find relevant cases*
   - formulate search queries to get all the causal connectives in the corpora
   - challenge: distinguish ‘false positives’ (e.g., *om* as preposition) from ‘true positives’.

How does it work?

3. *Identification of S_1 and S_2*
   - detect the size of the text segments connected by causal connectives automatically
   - challenge: incomplete S_1/S_2, embedded segments, constructions divided over different speakers (conversations, chat)

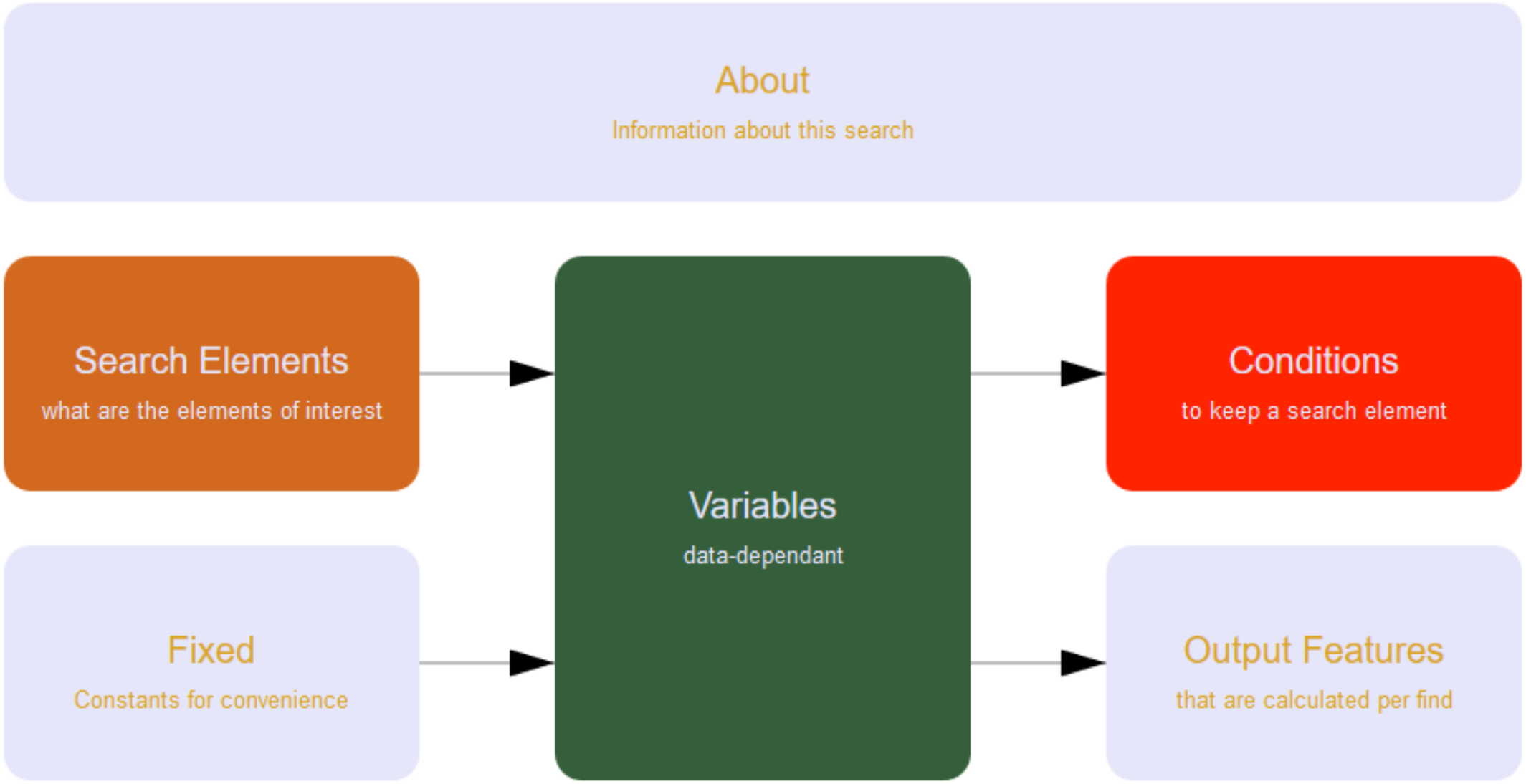
4. *Identification of the direction of the causal connection*
   - forward, backward?
   - depends on (i) type of connective (subordinating/coordinating conjunction, adverbs), (ii) semantics of the connective and (iii) grammatical position (“conn-S_1-S_2” versus “S_1-conn-S_2”)

How does it work?

5. Assessment of the subjectivity of the connection
   - how subjective is S_1/S_2? thematic text analysis (Bestgen et al., 2006)
   - the search interface determines features of the connection, such as the number of subjective adjectives and adverbs in S_1/S_2

ACAD: Automatic Coherence Analysis of Dutch

The search interface



ACAD: Automatic Coherence Analysis of Dutch

Editing the search: specification of variables

Search specification

Research project: **Want_Final?_2-E_2**

<< Overview

Show summary

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.

Provide the names and descriptions of the data-dependant variables here below. Make sure they are defined in the right **order**: any variable can only make use of other variables that are *above* it.

Name	Description		DELETE?
mrktype	Category of the marker in Dutch	Specify for search elements...	☐
markerCount	The number of words the marker consists of (1, 2, ...)	Specify for search elements...	☐
role	Direction into which causality is expected to go for this ma...	Specify for search elements...	☐
causativity	Kind of causativity (semantic/pragmatic)	Specify for search elements...	☐
blni	Boolean that indicates whether the marker is 'initial'	Specify for search elements...	☐
ini_type	Position of marker with respect to clause-start and to S1/S2	Specify for search elements...	☐
cpModFirstGuess	Determine whether the ancestor/parent of a marker plays a ro...	Specify for search elements...	☐
cpModParent	If the first guess of the CP has a REL or WH parent, take th...	Specify for search elements...	☐
cpMod	Any CP-type ancestor above the marker that must be taken int...	Specify for search elements...	☐
bCpIsVerbal	Any CP-type above the marker must contain a verb (WW)	Specify for search elements...	☐
blsPartOfPP	Is the parent of the marker a PP?	Specify for search elements...	☐
blsPartOfVC	Some markers may not be part of a VC - verbal complement	Specify for search elements...	☐
bConjPosOkay	Some markers must have a particular pos (e.g 'om' = VZ-CMP)	Specify for search elements...	☐
s1	List of word-nodes belonging to S1	Specify for search elements...	☐

ACAD: Automatic Coherence Analysis of Dutch

Controling the output

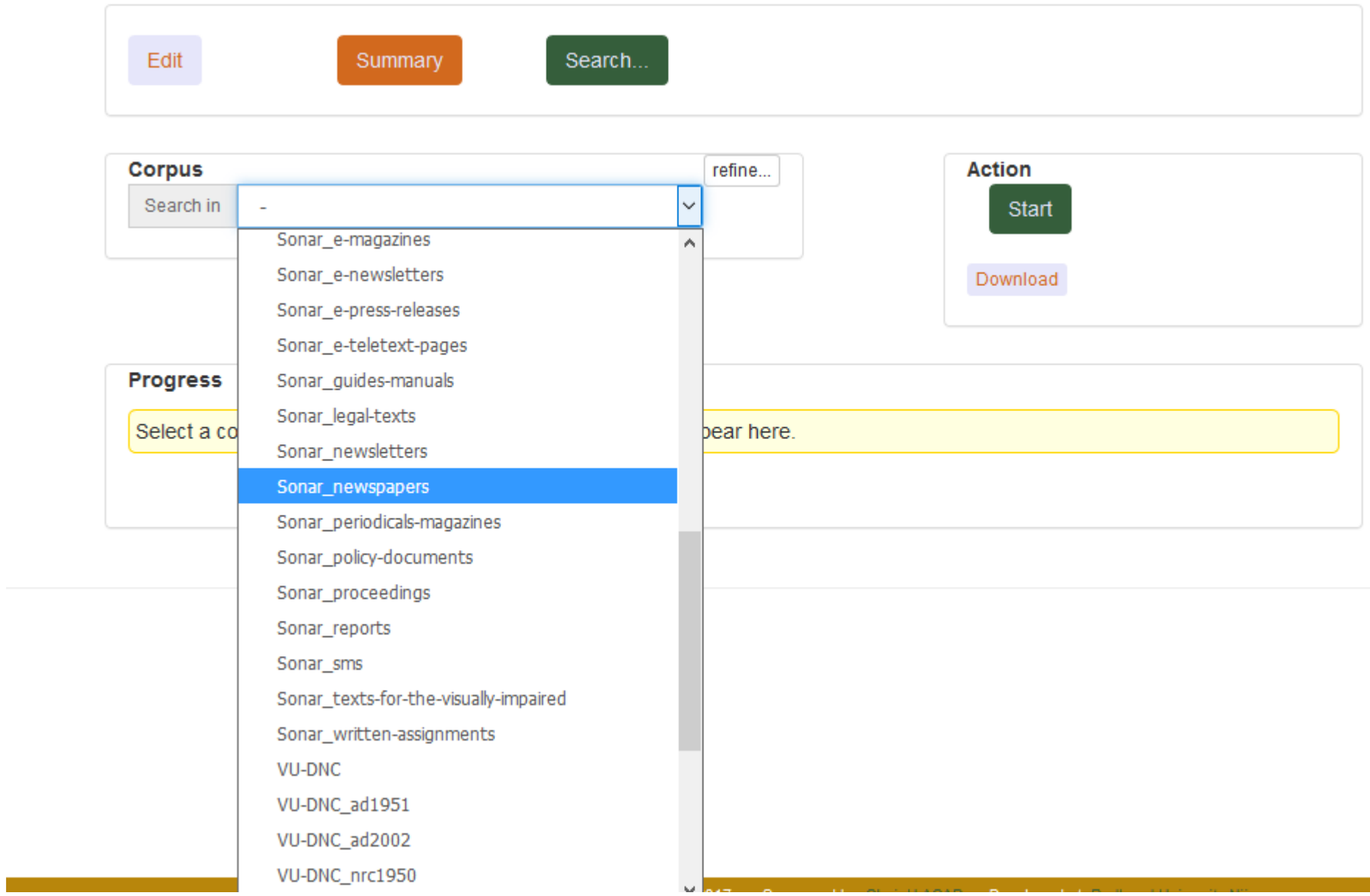
Output Features

that are calculated per find

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

Carrying out the search: choose the corpus



ACAD in action

Thematic text analysis

- Hypothesis
  - subjective connectives have more subjective words in their ‘consequent’ than objective connectives
    - backward causals: S1$_{want}$ > S1$_{omdat}$
    - forward causals: S2 $_{dus}$ > S2 $_{daarom}$

- Method
  - gold1000 lexicon of subjective adjectives (De Smedt & Daelemans, 2012)
    - 1044 adjectives, each word rated by seven raters
  - subjective: adjectives with a score of 0.7 or higher for each of their meanings (n=653); objective: adjectives with a score ≤ 0.2 (n=171)
  - the number of adverbial and adjectival uses of these adjectives was counted in both S_1 and S_2 for each of the four connectives

- Subcorpus: Sonar newspaper corpus (>700k news articles)

ACAD: Automatic Coherence Analysis of Dutch

Results

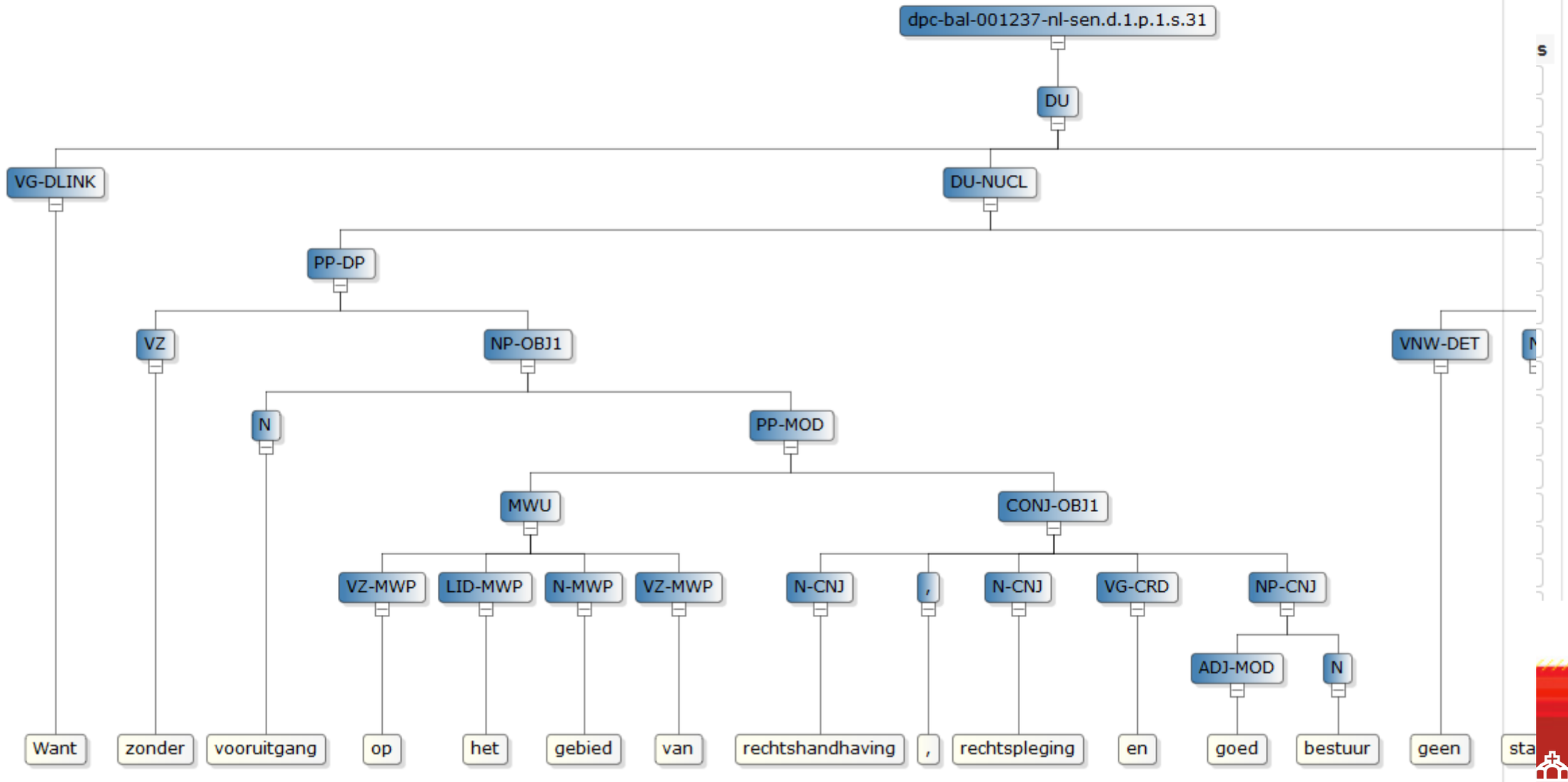
Search results

Search result line: 1 (472)

<< Result center

<< Back

Informatie...

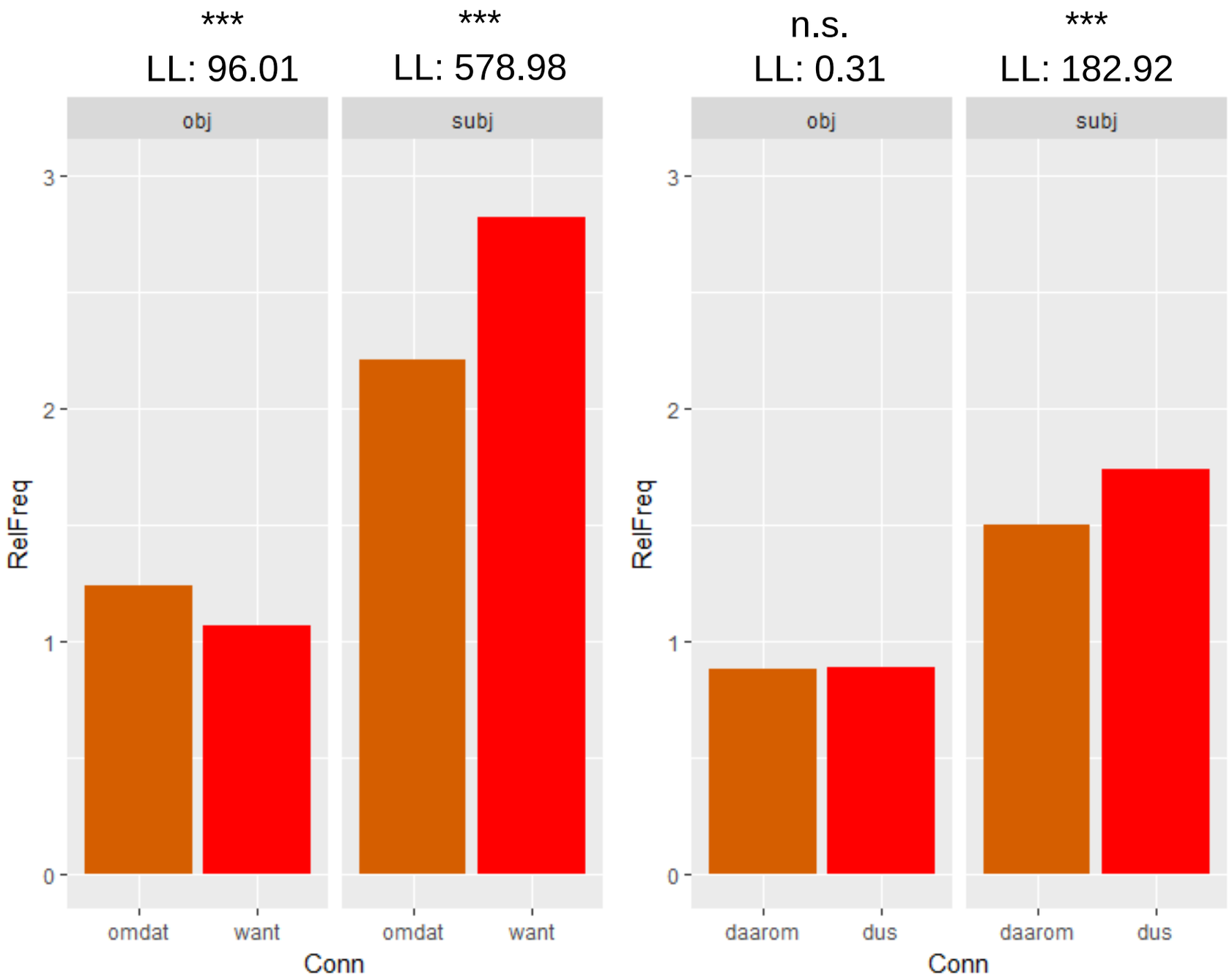


Results

	Subjective	Objective
Forward	Dus: 109,733	Daarom: 52,483
Backward	Want: 48,136	Omdat: 114,798

ACAD: Automatic Coherence Analysis of Dutch

Results



Next steps

- optimize precision and recall
- add examples of queries
- make manual

ACAD: Automatic Coherence Analysis of Dutch

SWOT analysis

- Potential
  - allows f
    searche
    – e.g., c
      of ‘sul
  - the grai
    be usec
  - queries
    adaptec
  - in princ
    any cor
    – provic
      forma
      tagged/lemmatizesd
- Challenges