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

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**)
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
Comparison 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?

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

![Relation type as function of connective and genre](image)
Using automatic analyses

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

ACAD: Automatic Coherence Analysis of Dutch
Goals of ACAD

- 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 reproducible (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
  - WhatsApp data of different age groups (13/14, 20-25)
ACAD: Automatic Coherence Analysis of Dutch

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’.
ACAD: Automatic Coherence Analysis of Dutch

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-$S_2$")
ACAD: Automatic Coherence Analysis of Dutch

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

About
Information about this search

Search Elements
what are the elements of interest

Fixed
Constants for convenience

Variables
data-dependent

Conditions
to keep a search element

Output Features
that are calculated per find
**ACAD: Automatic Coherence Analysis of Dutch**

**Editing the search: specification of variables**

---

**Search specification**

Research project: Want_Final?_2-E_2

---

The value of data-depdendant 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-depdendant 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.

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Specify for search elements</th>
<th>DELETE?</th>
</tr>
</thead>
<tbody>
<tr>
<td>mixinType</td>
<td>Category of the marker in Dutch</td>
<td></td>
<td></td>
</tr>
<tr>
<td>markerCount</td>
<td>The number of words the marker consists of (1, 2, ...)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>role</td>
<td>Direction into which causality is expected to go for this marker</td>
<td></td>
<td></td>
</tr>
<tr>
<td>causativity</td>
<td>Kind of causativity (semantic/pragmatic)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>bni</td>
<td>Boolean that indicates whether the marker is 'initial'</td>
<td></td>
<td></td>
</tr>
<tr>
<td>initype</td>
<td>Position of marker with respect to clause-start and to S1/S2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cpModFirstGuess</td>
<td>Determine whether the ancestor/parent of a marker plays a role</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cpModParent</td>
<td>If the first guess of the CP has a REL or WH parent, take this</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cpMod</td>
<td>Any CP-type ancestor above the marker that must be taken into account</td>
<td></td>
<td></td>
</tr>
<tr>
<td>bCOpasVerbal</td>
<td>Any CP-type above the marker must contain a verb (WW)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>blsPartOfPP</td>
<td>Is the parent of the marker a PP?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>blsPartOfVC</td>
<td>Some markers may not be part of a VP - verbal complement</td>
<td></td>
<td></td>
</tr>
<tr>
<td>bConfPosOkay</td>
<td>Some markers must have a particular pos (e.g. 'on' - V2-CMP)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>s1</td>
<td>List of word-nodes belonging to S1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## ACAD: Automatic Coherence Analysis of Dutch

### Controlling the output

<table>
<thead>
<tr>
<th></th>
<th>Description</th>
<th>Function</th>
<th>Args</th>
<th>Dependant</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>words in S1</td>
<td>Function</td>
<td></td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>last word in S1</td>
<td>Function</td>
<td></td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>list of word categories in S1</td>
<td>Function</td>
<td></td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>words in S2</td>
<td>Function</td>
<td></td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>whether S1 contains a verb</td>
<td>Data-depant variable</td>
<td>s1nodeVerbal</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>whether S2 contains a verb</td>
<td>Data-depant variable</td>
<td>s2nodeVerbal</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>category of copModAnc_cat</td>
<td>Function</td>
<td></td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>lemmas of neutral adjective in S1</td>
<td>Function</td>
<td></td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>number of neutral adjectives in S1</td>
<td>Function</td>
<td></td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>lemmas of subjective adjective/adverb in S1</td>
<td>Function</td>
<td></td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>number of subjective adjectives/adverbs in S1</td>
<td>Function</td>
<td></td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>number of words in S1</td>
<td>Function</td>
<td></td>
<td>yes</td>
<td></td>
</tr>
</tbody>
</table>
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: $S_1\text{want} > S_1\text{omdat}$
  - forward causals: $S_2\text{dus} > S_2\text{daarom}$

• 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
<table>
<thead>
<tr>
<th></th>
<th>Subjective</th>
<th>Objective</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forward</td>
<td>Dus: 109,733</td>
<td>Daarom: 52,483</td>
</tr>
<tr>
<td>Backward</td>
<td>Want: 48,136</td>
<td>Omdat: 114,798</td>
</tr>
</tbody>
</table>
ACAD: Automatic Coherence Analysis of Dutch Results

<table>
<thead>
<tr>
<th></th>
<th>obj</th>
<th>subj</th>
<th></th>
<th>obj</th>
<th>subj</th>
</tr>
</thead>
<tbody>
<tr>
<td>***</td>
<td>LL: 96.01</td>
<td></td>
<td>***</td>
<td>LL: 578.98</td>
<td></td>
</tr>
<tr>
<td>n.s.</td>
<td>LL: 0.31</td>
<td></td>
<td>***</td>
<td>LL: 182.92</td>
<td></td>
</tr>
</tbody>
</table>

Bar charts showing frequency of word usage in different contexts.
ACAD: Automatic Coherence Analysis of Dutch

Next steps

• optimize precision and recall
• add examples of queries
• make manual
ACAD: Automatic Coherence Analysis of Dutch SWOT analysis

• Potential
  - allows for sophisticated searches
    – e.g., other operationalisations of ‘subjectivity’
  - the grammatical parse can be used optimally
  - queries available that can be adapted
  - in principle extendable to any corpus and language
    – provided that it fits the Folia format and is parsed/pos-tagged/lemmatized

• Challenges
  - the system is not as flexible and stable as we would like
  - queries on large corpora can take long
  - formulating the queries can be challenging