CLARIN project "DiscAn":

Towards a Discourse Annotation system for Dutch language corpora

Ted Sanders
Kirsten Vis
Utrecht Institute of Linguistics
Utrecht University

Daan Broeder
TLA – Max-Planck Institute for Psycholinguistics Nijmegen

Universiteit Utrecht

Discourse coherence in annotated corpora ?

- international tendency towards discourse annotation:
  - Penn Discourse Treebank (Prasad, Joshi, Webber et al.)
  - Potsdam Corpus (Stede et al.)

- in annotated Dutch corpora this discourse level is largely lacking
- but at the same time, much data on Dutch
- in our case:
  - on connectives
  - mainly causal
  - across media (various written genres, spoken, chat)
  - at various stages of annotation

Overview

1. curation of annoted corpora
    - overview of data
    - Example

2. development of discourse annotation system
    - minimal set of characteristics
    - relation to existing work (PDTB; Project ISO 24617-8: Semantic Relations in Discourse)
    - example of annotated case

3. discussion of relation to existing work

Goal 1. Curation

Discourse phenomena	Author	Cases
Causal connectives	Bekker (2006)	500 (*doordat, want, dus, daarom, nadat, voordat*) / 200 implicit
Causal connectives	Degand (2001)	150 *(want, aangezien, omdat)* from newspapers
Coherence relations, discourse structure	Den Ouden (2004)	70 (causal implicit, non-causal)
Connectives	Evers-Vermeul (2005); Stukker (2005)	600 historical data (*want, omdat, dus, daarom*) / 4400 from Childes (*en, maar, toen, want*)
Causal connectives	Pander Maat & Degand (2001)	150 (*dus, daarom*) from newspaper corpora
Coherence relations	Pander Maat & Den Ouden (2011)	795 implicit and explicit relations from a self-assembled corpus of 40 press releases
Causal connectives	Pander Maat & Sanders (2000)	150 (*dus, daarom, daardoor*) from a newspaper-corpus (*Volkskrant*)
Causal connectives	Persoon (2010)	105 (*omdat, want*) from CGN
Causal connectives	Pit (2003)	200 (*aangezien, omdat, doordat, want*) newspaper / 100 (*omdat, doordat, want*) narrative; from newspaper (*Volkskrant*) and fictional books
Causal connectives	Sanders & Spooren (2009); Spooren et al. (2010)	100 newspaper (*Volkskrant*) / 275 from CGN / 80 from Chat (*want, omdat*)

Curation (2)

Discourse phenomena	Author	Cases
Coherence relations and connectives	Spooren & Sanders (2008)	1100 coherence relations (children elicit responses)
Causal connectives	Stukker (2005)	300 (*daardoor, daarom, dus*) newspaper / 300 historical data (*daarom, dus*)
Coherence relations	Vis (2011)	135 texts; 643 subjective relations

Corpora

Diversity in available data:

- type of discourse:
  - written (newspaper, fiction)
  - spoken (CGN, child language)
  - chat
- 'new corpus' or existing corpus
  - (annotation layers in some existing corpora)
- complete text or pairs of segments
- annotation of connectives y/n
- different annotation models
- different formats (Word/txt/SPSS/Excel/XML)

Example from spoken language

Corpus: Persoon (2010); 105 cases from Corpus of Spoken Dutch (CGN)

dan bijvoorbeeld dat meisje wil misschien dan als ze in mijn kamer heeft gezeten dan misschien jouw kamer misschien wel weer overnemen maar dat moeten wij eigenlijk nog helemaal niet zeggen {want} misschien vinden wij we die helemaal helemaal niet leuk

then for example that girl might want when she has been in my room then maybe she wants to maybe take over your room but we should probably not say that yet {because} maybe we we don't like her at all at all

[example WANTHZ35 from file fn000683 (face-to-face)]

Analysis (1)

Characteristic	Value
nr	35
type of causal relation	speech-act
modality X	intentional action
modality Y	evaluation
cp	implicit
role cp	speech-act
expression cp	implicit
nature cp	speaker
perspective	speaker
passive X	no
passive Y	no

Analysis (2)

Characteristic (2)	Value (2)
tense X	ott
tense Y	ott
polar elements X	maar, niet
polar elements Y	niet
modal elements X	moeten, eigenlijk
type of modality	deontic, epistemic
modal elements Y	misschien
type of modality	epistemic
size segment X	clause
size segment Y	clause
form segment X	directive
form segment Y	confirmation
position segment X	direct
speaker continuity	same speaker
position connective	start Y
syntactic modification	none
argumentation structure	singular

Curation of files

- standardize corpus texts
  - collect texts and analyses
  - convert to XML (PAULA format; *Potsdamer Austauschformat für linguistische Annotation*, "Potsdam Interchange Format for Linguistic Annotation")
  - unify annotations

    → develop discourse annotation system

Goal 2. Development discourse annotation system

- Goal: standardize annotation AND develop system for future annotations

- proposal: MINIMAL SET of characteristics
  - 3 types: administrative, relational and re segments
  - systematic: cross-classification defines categories
  - express 'family resemblance'
  - fundamental characteristics, present in all proposals
  - other characteristics can be derived, so compatible to most proposals

- Relational characteristics
  - Polarity — positive | negative
  - Basic operation — causal | additive | temporal
  - Source of coherence — content | epistemic | speech-act | textual
    - if Source of coherence = content, Volitionality — volitional | non-volitional
  - Order — forward (S1=P, s2 = Q) | backward (S1=Q, S2 = P)

  - Linguistic marking — yes | no
  - Connective (or other lexical marker) — aangezien | daardoor | daarom | doordat | dus | omdat | want | etc.

What is this minimal set based on?

- Theories of relations: taxonomies
  (Martin92, Sanders et al 92,93, Kehler 2002; MannThompson 88)

- Corpus work on connectives
  (Degand, Knott, Pander Maat, Sanders, Spooren, Sweetser, ….):
  how is the lexicon of connectives of various languages organized?
  - all languages distinguish causal – temporal – additive
  - only Dutch distinguishes volitional vs. Non- volitional

- Cognitive considerations: relevance of these categories
  - acquisition: positive before negative; additive < temporal < causal
  - processing: causals faster than additives; epistemic slower than content
  - representation: Causals recalled better than additives and temporals.

Minimal set

- possible other advantages
  - make decisions step by step
  - training of annotators
  - use Paraphrase and Substitution tests

  - prediction: easier
  - more reliable?
  - future work: reliability experiment

- relation to other proposals
  - PDTB; Project ISO 24617-8: Semantic Relations in Discourse
  - more systematic
  - BUT: compatible with PDTB

Compatibility with PDTB

- Third-quarter sales in Europe were exceptionally strong, boosted by promotional programs and new products although weaker foreign currencies reduced the company's earnings.
  (PDTB; Contingency:concession:contra-expectation)

  Minimal set:
  - Polarity negative
  - Basic operation causal
  - Source of coherence content
    - Volitionality non-volitional
  - Order backward

  - Linguistic marking yes
  - Connective although

Compatibility with PDTB

- She became an abortionist accidentally, and continued <u>because</u> it enabled her to buy jam, cocoa and other war-rationed goodies.
  (PDTB; Contingency:cause:reason)

  Minimal set:
  - Polarity positive
  - Basic operation causal
  - Source of coherence content
    - Volitionality volitional
  - Order backward

  - Linguistic marking yes
  - Connective because

Minimal set: characteristics of segments

- Characteristics of segments
  - modality of S1 — fact | situation (knowledge / experience) | judgment | intentional action
  - modality of S2 — fact | situation (knowledge / experience) | judgment | intentional action
  - Subject of Consciousness S1 — speaker-writer | 2nd person | 3rd person | generic 3rd person | secundary speaker / writer | not applicable
  - Subject of Consciousness S2 — speaker-writer | 2nd person | 3rd person | generic 3rd person | secundary speaker / writer | not applicable

Minimal set: administrative features

- Administrative features
  - corpus [string]
  - fragmentid [string]
  - sentence/line/id [string]
  - annotator Bekker | Degand | Degand & Pander Maat | Den Ouden | Evers-Vermeul | Pander Maat & Den Ouden | Pander Maat & Sanders | etc.

Uniform format

- Current:
  - fragments — Word / txt / XML
  - source texts — Word / txt / XML
  - analyses — SPSS / Excel / Word / XML
- In this project:
  - PAULA format; *Potsdamer Austauschformat für linguistische Annotation*, "Potsdam Interchange Format for Linguistic Annotation"
- Reasons:
  - web-based; architecture for querying, etc. (ANNIS2)
  - stand-off annotation
  - several layers of annotation possible
  - possibility of visualization (tree structure / segment pairs)

Conversion

Convert files

1. change to PAULA format

Convert analyses

1. recode analyses into minimal set
2. complete missing values
3. convert to PAULA format

Example of annotated case

Source text:

SandersSpooren_fragm3505.text.xml

```
1  <?xml version="1.0" encoding="UTF-8"?>
2  <header paula_id="corpusSandersSpooren_3505_text" type="text"/>
3  <body>
4  Uit elkaar
5  Actrice Julia Roberts en haar vriend Benjamin Bratt hebben een punt gezet achter hur
6  Ze wilde niet trouwen en Bratt zou daarom hebben aangestuurd op het verbreken van de
7  </body>
8
```

Analysis:

```
<featList type="discrel" xml:base="SandersSpooren_fragm3505.text.xml"
   <feat xlink:href="#tok44" value
   corpus=SandersSpooren; fragmentid=3505 uitingid=910;
   annotator=SandersSpooren; modality of S1= judgment; modality of
   S2=judgment; SubjofConscS1= 31; SubjofConscS2=31; polarity=? order=?
   basic operation=3; source of coherence=epistemic;linguistic
   marking=yes; connective=omdat;
```

Analytical decisions needed to arrive at a discourse annotation of coherence relations

0. Determine S1 and S2: usually clauses; does the relation hold between adjacent segments?

1. Basic operation: P & Q or P -> Q ?

2. Polarity: P and Q or negation(s) ?

3. Source of Coherence:
   - Objective / content: two situations / facts / events / locutions
   - Subjective / epistemic or speech act: illocution "the saying of" or speaker conclusion involved

4. Order of the segments: Forward / Backward

Tests for the analysis of coherence relations

Paraphrase tests

- Basic operation:
  "and also" or "and then" versus "this leads to" or "is caused by"

Within causals:

- Objective / content:
  - "Situation p leads to Situation q"
  - "Situation q is caused by Situation p"
- Subjective /epistemic
  - "Situation p leads to my / speaker's conclusion q"
  - q is my / speaker's claim based on argument p

Tests for the analysis of coherence relations

Paraphrase tests

- subjective / speech act
  - "Situation p leads to my / speaker's saying q."
  - "My / Speaker's saying of q is caused by situation p".

Substitution tests

- When relation is implicit: which connective / signal expresses the relation best?