



# Building a Discourse annotated corpus for Dutch

Daan Broeder (TLA/MPI),  
Kirsten Vis (UU),  
Ted Sanders (UU)

# A Dutch discourse annotation corpus

---



## Project objectives:

- Curate and make available an existing set of Dutch corpus analysis of coherence relations and discourse connectives
- Develop a discourse annotation system applicable for Dutch
- Contribute to the CLARIN effort by developing metadata profiles for this type of data
- Inventory of discourse categories and sync with the ISOcat discourse categories effort
- Discourse community building in NL and B

# CLARIN Curation Projects

---



- Purpose
  - Making (existing) data and tools available in a sustainable way
- Strategic choices
  - Covering of linguistic and SSH disciplines
- CLARIN requirements: organizational & technical
  - Stable center
    - Who to phone when things go wrong
    - Expectation of persistence
  - Resource discovery using CMDI
  - Sustainable access
    - Persistent Identifiers for resources
    - IPR issues solved or at made least explicit
  - Explicit semantics by describing (new) concepts in the ISOcat concept registry.
    - For specific metadata elements
    - ... and also annotation labels

# The Language Archive

---



- TLA is a unit of the MPI for Psycholinguistics
  - Funded by MPG, MPI-PL, KNAW, BBAW
  - Home of the institute research corpora: L&C, Acquisition, Sign Language, ...
  - Houses one of the largest Endangered Languages documentation archives
  - Archive for several Dutch & German corpora where projects have no proper archive facilities
  - Offer for researchers to deposit “serious” data
- Software development
  - Archiving software
  - Exploitation tools: ELAN, ...
  - Research infrastructure projects: CLARIN, DASISH, EUDAT
- CLARIN center status: B center, A center status pending
- TLA ambitions
  - More data: linguistic domain, wider humanities.
  - Build relations with (new) communities & research groups and come to common project proposals
  - Discourse annotation an opportunity to broaden our basis and
  - ... gain expertise with Discourse Annotation specific needs (tools)

# Available data for curation

---



- Results from 16 discourse research projects analysing coherence relations and connectives
- Formats:
  - Fragments from newspaper articles, Childes or other corpora (CGN)
  - Excel, MS-Word, plain text, ...
- Metadata available in files of varying formats and from publications
  - Corpora: structured metadata
  - Newspapers: none or not structured
- Varying annotation schemes used
- IPR issues:
  - Childes -> enrichments should be offered to community
  - NLTU © corpora (CGN, D-COI) -> enrichments also available via TST-centrale
  - Newspaper sources: many fall within existing agreements for research use other need to be negotiated

# Corpora (1)



Author	phenomena	Cases	Source	Remarks
Pander Maat & Sanders (2000)	Causal connectives	150 (dus, daarom, daardoor)	- Newspapers	
Degand (2001)	Causal connectives	150 (want, aangezien, omdat)	- Newspapers	
Pit (2003)	Causal connectives	- 200 (aangezien, omdat, doordat, want) - 100 (omdat, doordat, want) narrative	- Newspapers - Newspapers and fictional books	Split into two distinct subcorpora, different metadata sources

# Corpora (2)



Author	phenomena	Cases	Source	Remarks
Stukker (2005)	Causal connectives	<ul style="list-style-type: none"><li>- 300 (daardoor, daarom, dus)</li><li>- 300 (daarom, dus)</li></ul>	<ul style="list-style-type: none"><li>- Newspapers</li><li>- Historical data</li></ul>	Split into three structurally similar but distinct subcorpora: daardoor, daarom, dus (mixed sources)
Sanders & Spooren (2009)	Causal connectives	<ul style="list-style-type: none"><li>- 100</li><li>- 275</li><li>- 80 (want, omdat)</li></ul>	<ul style="list-style-type: none"><li>- Newspapers</li><li>- CGN</li><li>- Chat</li></ul>	Relatively little metadata available (sources are not publications)
Persoon (2010)	Causal connectives	<ul style="list-style-type: none"><li>- 105 (omdat, want)</li></ul>	<ul style="list-style-type: none"><li>- CGN</li></ul>	

Planned curation data reduced from 7000 cases to < 2000

# The Creation of a Corpus

---



- Original material was highly diverse
  - Origin: newspapers, annotated speech corpora, ..
  - Formats: SPSS, MS-Word, ...
  - Annotations: specificity, references to primary data
- Curation required:
  - Consistency in formats, annotation labels, metadata
  - Metadata format determined by CLARIN requirements -> CMDI
  - Annotation format determined by the exploitation environment
- Limited time and resources forced some choices
  - Difficult to plan, surprises will happen
  - 16 data-sets -> 7 fully curated data-sets -> 9 corpora
  - Depth or breadth? Depth!
- Other corpora can be handled now much more efficiently



# Exploitation Environment

---



- Archiving alone is not sufficient
- Exploitation environment is required
- Existing TLA/MPI tools are not sufficient
  - Metadata search & (limited) content search
  - Not directed to discourse annotations
- Need appropriate domain specific search & visualization

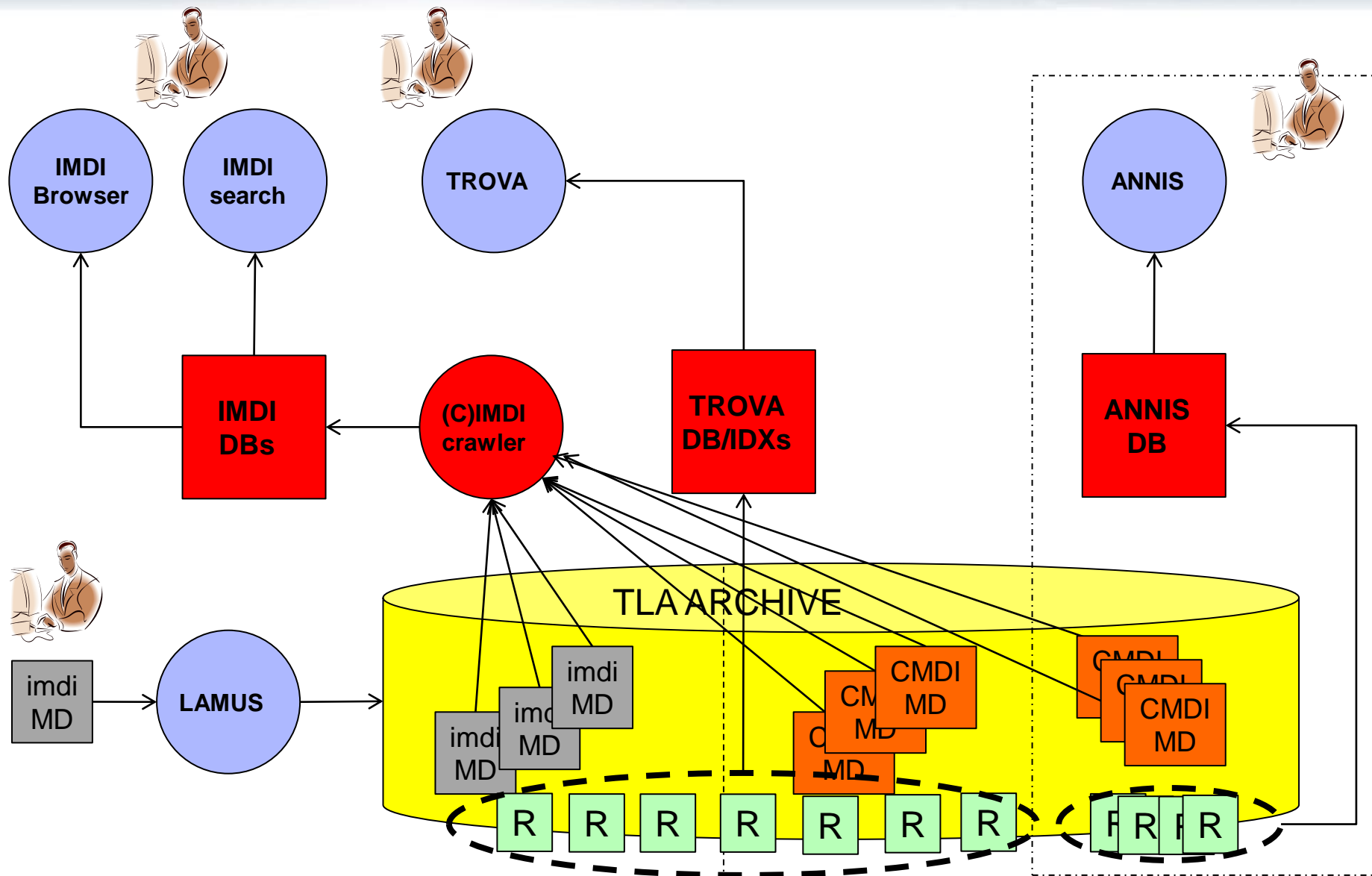
# Tools considered

---



- Penn Discourse Treebank
  - Format considered problematic
  - Not a web application
- ANNIS
  - Sustainability: fast support, docs, Java, RDB,
  - Scalability? For the moment not needed
  - Search GUI is ok, regexp support, SQL only
  - Visualization is ok.
  - Possibility for close collaboration
- Build/Extend own software
  - Not if something suitable is available
  - Not in this project with limited resources

# TLA Archive: LAT & ANNIS



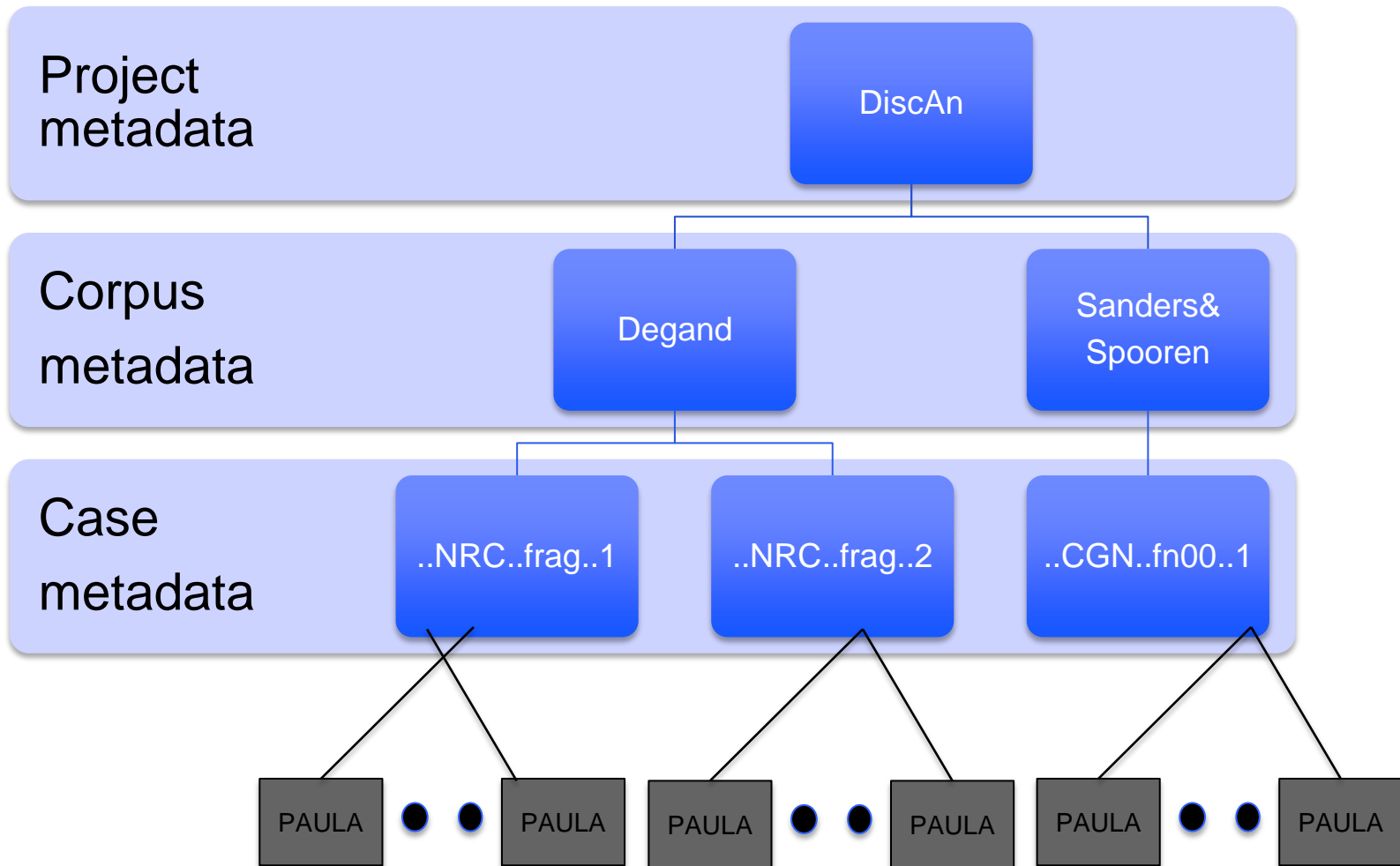
# Metadata creation

---



- Different types of metadata sources, all corpus specific, sometimes sub corpus specific
  - Extraction from PAULA files, (sub) corpus specific mappings
  - Extraction from existing (hand crafted) tables
- Curation of metadata records using a specialized CMDI metadata editor (ARBIL)
  - Allows table manipulation
  - Search over metadata file sets

# Corpus structure



# (C)IMDI view



The screenshot shows a web browser window with the URL `corpus1.mpi.nl/ds/imdi_browser/?openpath=MPI1708763%23`. The page title is "IMDI-Browser". At the top right, there are links for "hide accessibility of resources", "about", "manual", "register", "user: anonymous", "login", and "logout".

The left sidebar shows a tree view of corpora under "IMDI-corpora":

- index.html
- WelcomeToIMDIDomain.html[1]
- AILLA
- ANDES
- CLARIN NL
  - DiscAn
    - Degand22
    - PanderMaatSanders
    - Persoon
    - Pit\_narrdow**
      - Pit AKK05 [0]
      - Pit AKK10 [0]
      - Pit AKK16 [0]
      - Pit AKK80 [0]
      - Pit BEC08 [0]
      - Pit BEC14 [0]
      - Pit BEC21 [0]
      - Pit BEC28 [0]
      - Pit BEC32 [0]
      - Pit BEC38 [0]
      - Pit DIS104 [0]
      - Pit DIS11 [0]
      - Pit DIS28 [0]
      - Pit DIS29 [0]
      - Pit DIS47 [0]
      - Pit DIS48 [0]
      - Pit DOR08 [0]
      - Pit DOR08 [0]
      - Pit DOR24 [0]
      - Pit DOR25 ...
      - Pit DOR43 [0]
      - Pit DOR91 [0]
      - Pit GIL07 [0]

The main content area is titled "Imdi Information" and "ISLE Metadata Initiative". It displays the following information for the selected corpus:

- Corpus**
  - Name** Pit\_narrdow
  - Title** DiscAn: Pit\_narrative corpus of doordat, omdat and want
- Description**

The Pit narrdow subcorpus was compiled for a study of the causal connectives 'aangezien', 'doordat', 'omdat' and 'want' in Dutch, German and French narratives and news. It consists of cases from 22 Dutch novels published between 1990 and 1996.
- Location**
- Project** Pit
- Keys**
  - Publication.Description**

Pit, M. (2003). How to express yourself with a causal connective? Subjectivity and causal connectives in Dutch, German and French. Amsterdam: Editions Rodopi B.V.
- Content**
- Actors**
  - Actor** Dr. Mirna Pit

[http://corpus1.mpi.nl/ds/imdi\\_browser?openpath=MPI1705091](http://corpus1.mpi.nl/ds/imdi_browser?openpath=MPI1705091)

# ANNIS view



IMDI Browser x Annis<sup>2</sup> Corpus Search x

https://lux16.mpi.nl/big-ds/annis/search.html

ANNIS<sup>2</sup> Tutorial logged in as "test"

### Search Form

AnnisQL: /twintig/

Show Result Query Builder History

Result: 8

More Corpora

<input type="checkbox"/>	Name	Texts	Tokens	
<input type="checkbox"/>	Degand22.txt	143	7316	
<input type="checkbox"/>	PanderMaatSanders.txt	100	15063	
<input type="checkbox"/>	Persoon.txt	105	3135	
<input type="checkbox"/>	Pit_narrow.txt	107	27422	
<input checked="" type="checkbox"/>	Pit_totadow.txt	198	64225	
<input type="checkbox"/>	SanderSpooren.txt	483	81639	
<input type="checkbox"/>	Stukker_daardoor.txt	94	9382	
<input type="checkbox"/>	Stukker_daarom.txt	94	10414	

Search Export

Context Left: 5

Context Right: 5

Results Per Page: 10

### Search Result - /twintig/ (5, 5)

Page 1 of 1 Token Annotations Show Citation URL Document Path

Displaying Results 1 - 8 of 8

wordt dat ? ? ? op de **twintig** beginnende high-technoedemeringen in Californi<sup>▼</sup> uiteindelijk

- Pit\_158 (tree)
- Pit\_158 (grid)
- paula
- paula text

Path: Pit\_totadow.txt > Pit\_173

tijdens een persoonlijk onderhoud dat **twintig** minuten duurde. Sinn Fein-leider Gerry

- Pit\_173 (grid)

Select Displayed Annotation Levels

tok tijdens een persoonlijk onderhoud dat **twintig** minuten duurde. Sinn Fein-leider Gerry

- Pit\_173 (tree)
- paula
- paula text

Path: Pit\_totadow.txt > Pit\_195

wil vandaag voorstellen dat er **twintig** proefboerderijen worden ingericht als voorbeeldbedrijven

- Pit\_195 (tree)
- Pit\_195 (grid)
- paula
- paula text

Path: Pit\_totadow.txt > Pit\_34

een traject afleggen van maximaal **twintig** zeemijl (ruim 37 kilometer). Lange

- Pit\_34 (tree)
- Pit\_34 (grid)
- paula
- paula text

Path: Pit\_totadow.txt > Pit\_66

een oerwoudtaal die nog door **twintig** indianen wordt gesproken. Haast is

- Pit\_66 (tree)
- Pit\_66 (grid)
- paula
- paula text

Path: Pit\_totadow.txt > Pit\_66

<https://lux17.mpi.nl/annis/search/search.html>

# Some problems encountered

---



- Organizational ones: two teams with different experience and perspective
  - UU team: curating the annotation data creating the metadata
  - TLA team: archiving and infrastructure part + mentoring UU-team wrt. requirements
  - UU-team with limited resources achieved good results
  - However misconceptions forced some redistribution of tasks
  - Documentation and time-period needs for transfer of knowledge and feedback were underestimated
- Technical problems
  - Encoding problems introduced by Python scripts
  - Soft-hyphens (MSWord) not well taken care of



# Future steps

---



- Specific for this project/corpus
  - Allow easy download whole corpus data-sets. Some may have a local ANNIS installation
  - Proper license for access to the data, if not free!!!
  - Start ANNIS from the general TLA catalog application
  - Index PAULA files also with the TLA indexer
- General
  - Tighter integration ANNIS with TLA software



---

Thank You for Your Attention



- 
- Maarten Postma
  - Jeroen Breteler
  - Twan Goosen
  - Eric Auer

# Archive Access



Browsing/Search/Visualization

WWW browser



Web apps.

**TLA ARCHIVE**

metadata

HTTP server

annotations

media files

Local tools



resource download

LAMUS

**LOCAL DATA**

All resources accessible by HTTP if authorized

