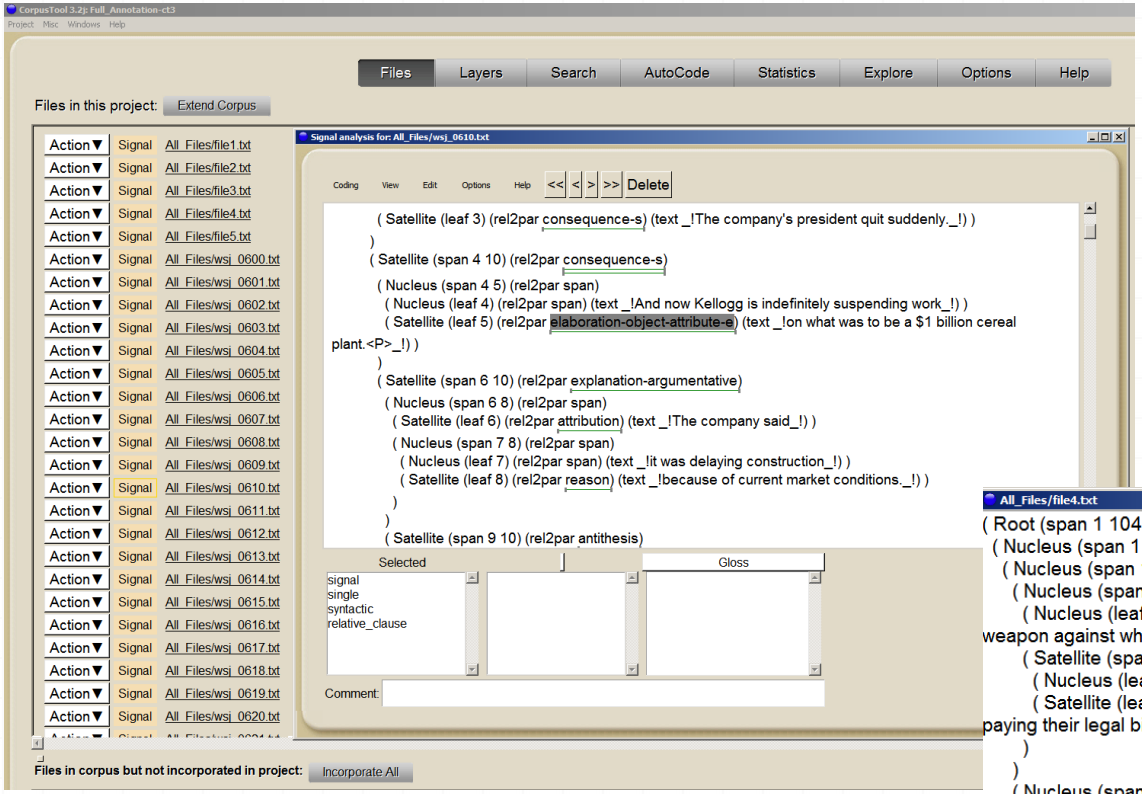
Discourse Annotation Tools overview

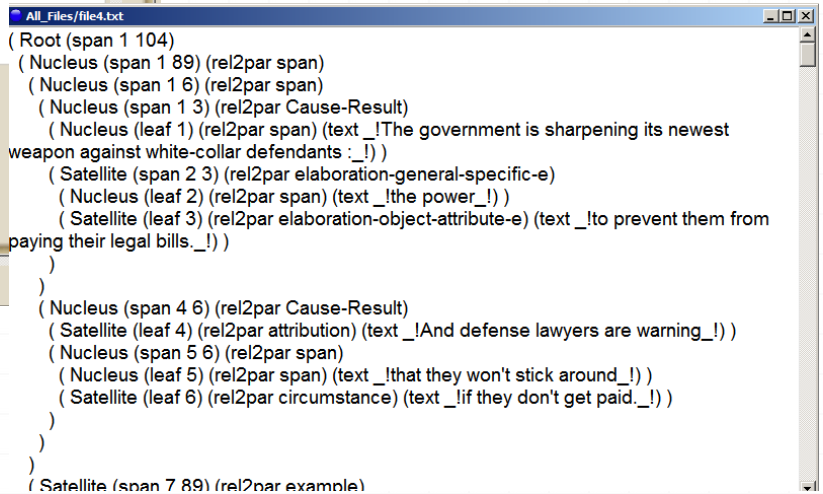
Anja Nedoluzhko

Discourse Annotation Tools

- free ↔ commercial (very many of all)
- created for annotation of spoken data ↔ written data ↔ general (different functions)
- widely accepted and used (WebAnno, brat, MMAX, EXMARaLDA, etc.) ↔ not so popular (often made for specific task)

UAM Corpus Tool3





RST Signalling Corpus (Das – Taboada, 2015)

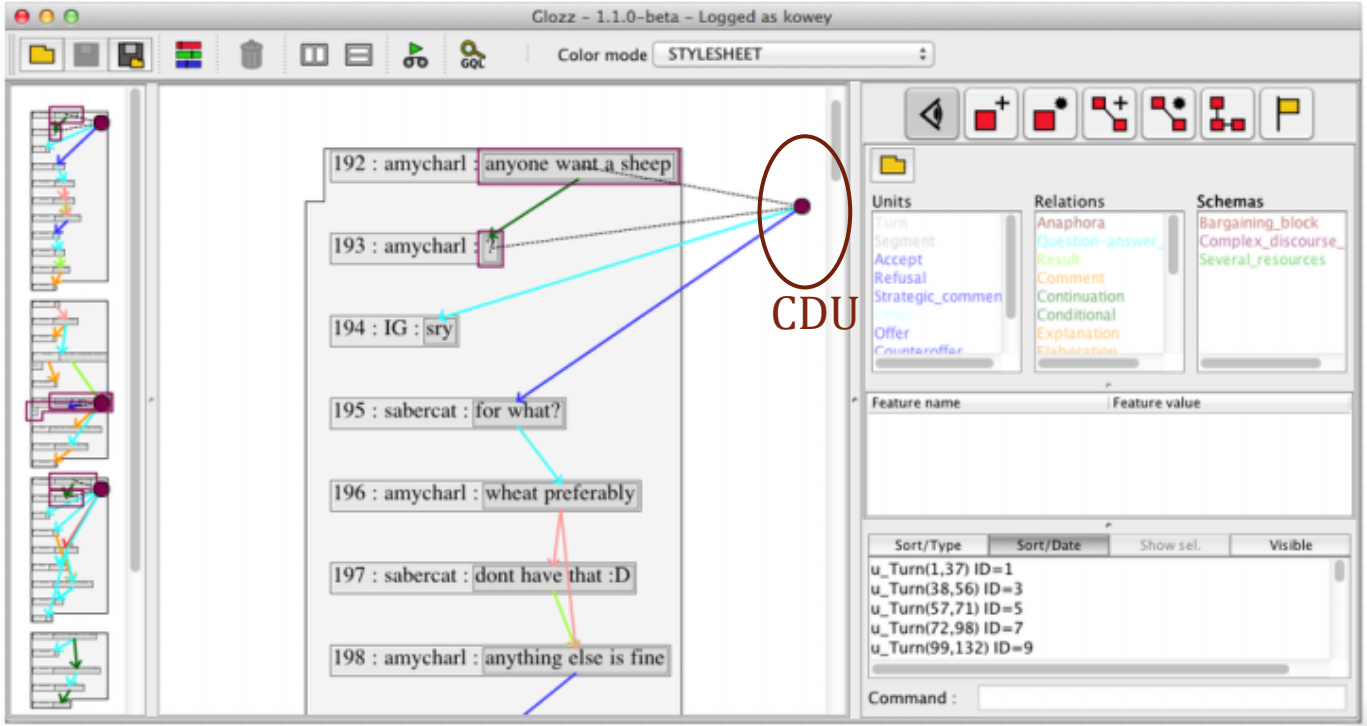
Tools for general aims

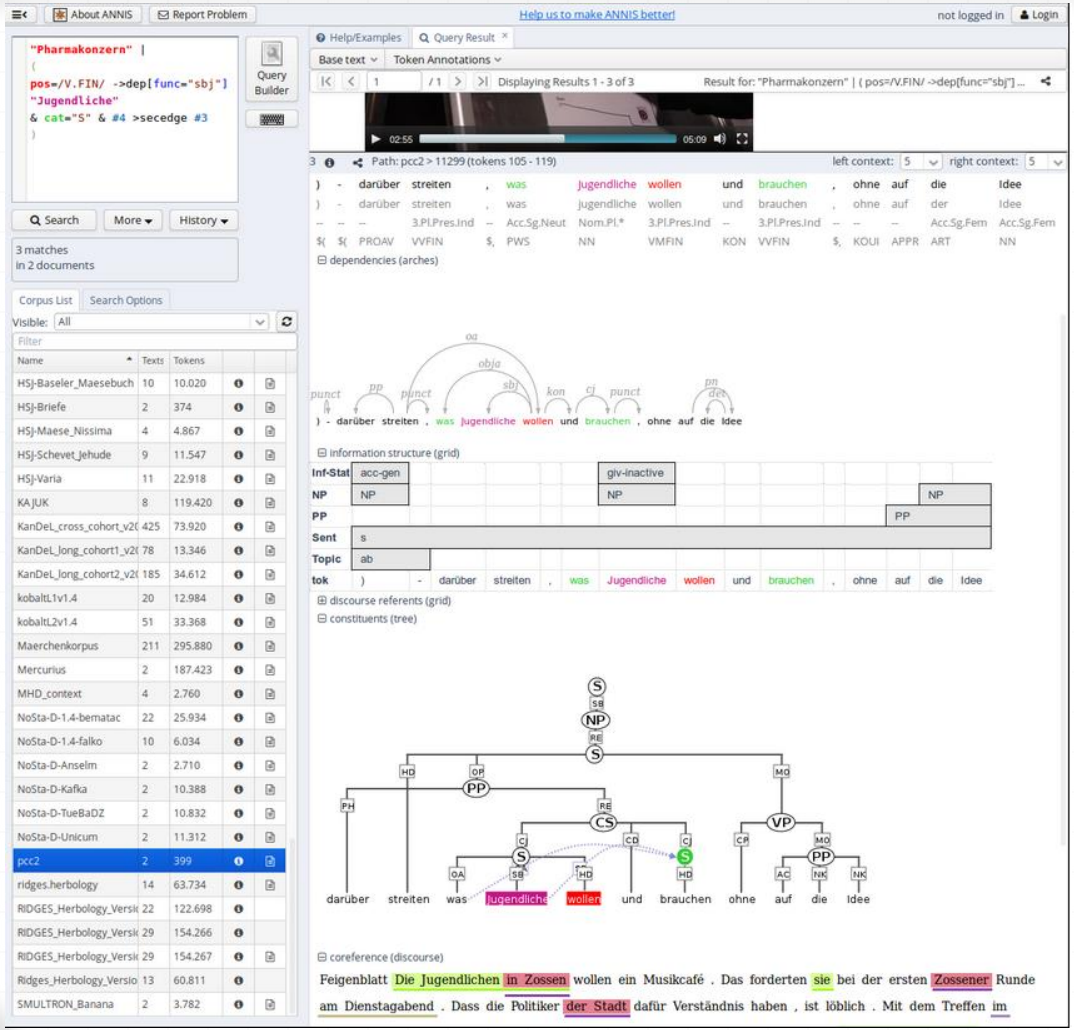
standalone

Glozz

used e.g. in STAC (Asher et al.)

- Developed in an SDRT annotation context
- Used e.g. in STAC (Asher et al.)
- It is able to point to annotations of any type, making it so that you can have eg. relations between schemas and units, relations and relations, etc.





ANNIS

- http://corpus-tools.org/annis/
- used for DiscAn annotations
- open source, cross platform (Linux, Mac, Windows), web browser-based search and visualization architecture

(ideal) PLAN for today

- short overview
- demonstration of **ELAN** and **MMAX**
- lab itself (annotation of wsj_2395):
  - demonstration of the annotation in **TrEd** (according to PDTB-like Prague annotation rules)
  - annotation in the **PDTB** annotation tool (according to PDTB-3 annotation rules)
  - annotation in the **RSTweb** tool (according to the RST theory)
  - annotation in the **brat** tool (according to CCR theory)

Tools – classification criteria

- general
- spoken
- written

- standalone
- server

- advantages
- disadvantages

WebAnno

brat LAB

MMAX DEMO

TrEd DEMO

EXMARaLDA

Praat

ELAN DEMO

PDTB LAB

RSTweb LAB

Tools concerned today

Tools for general aims

server and standalone, web based GUI

WebAnno

https://webanno.github.io/webanno/

ADVANTAGES:

- import different formats
- flexibility to define own annotation schema –
- annotation layers predefined for: lemma, POS, NE, coreference, dependencies...
- documentation (with videos)
- under active development
- responsive community
- good user interface
- machine learning module to try to learn a system from manual annotation
- one can keep the original sentence splitting and tokenisation of the text
- output supported in different formats
- multilayer annotation

Tools for general aims

server and standalone, web based GUI

WebAnno

https://webanno.github.io/webanno/

ADVANTAGES:

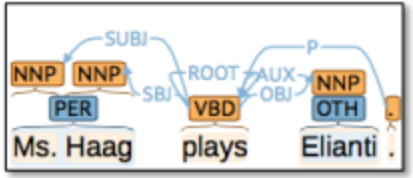
If installed in a server one can get the most of:
- managing projects with multiple annotators
- curation/revision tasks
- interannotator agreement
- monitoring progress of the project

DISADVANTAGES:

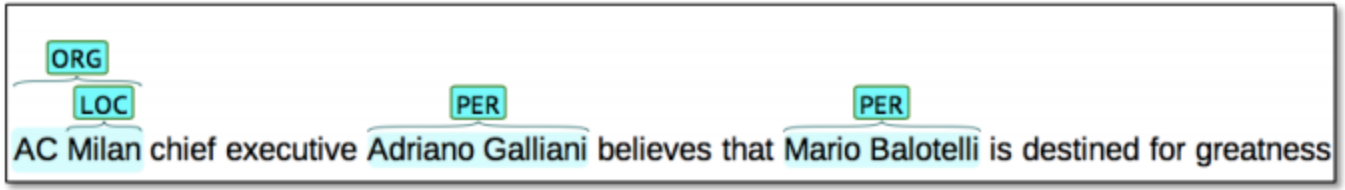
not possible or not easy to annotate some types of information: for example, constituents

WebAnno

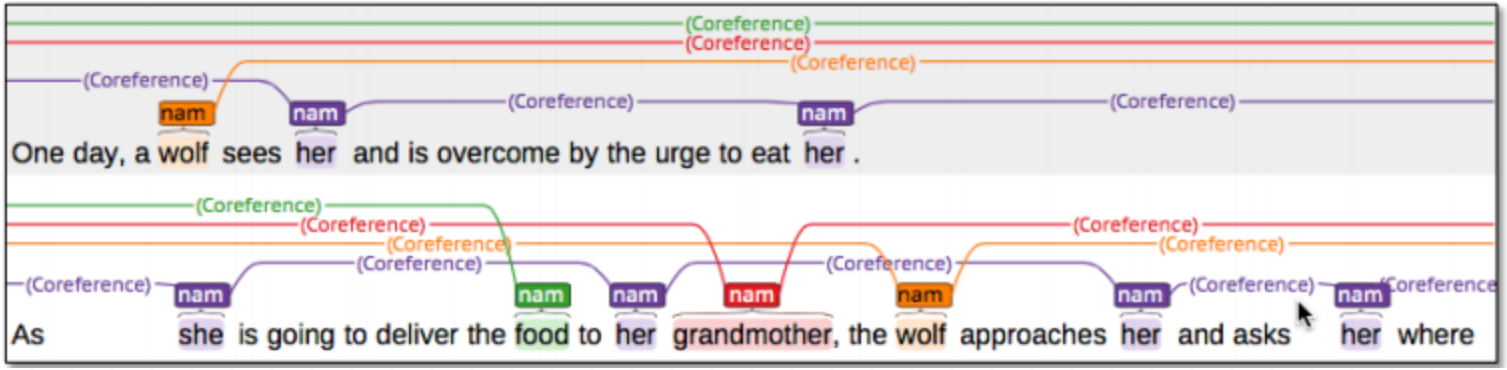
Part-of-Speech & syntactic dependencies



Named entities

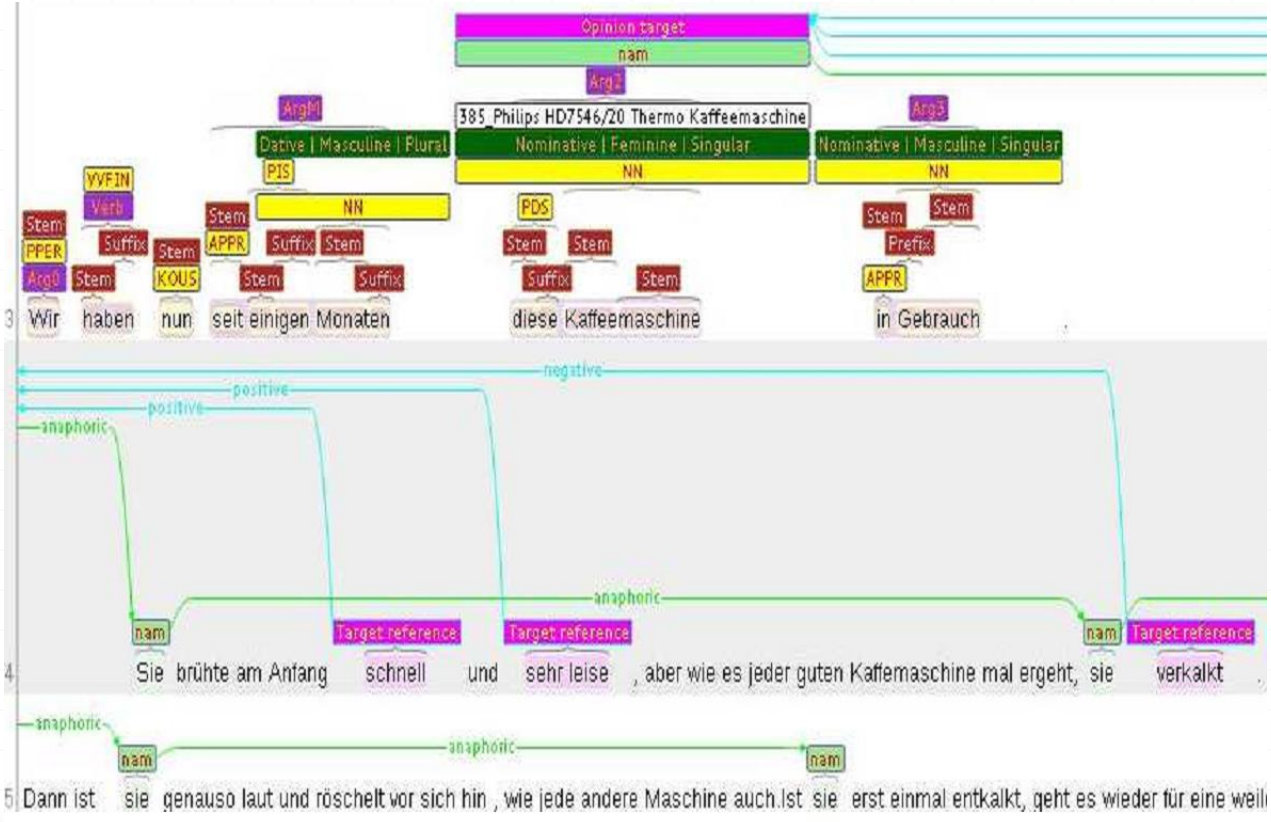


Co-reference



Annotation with multiple layers

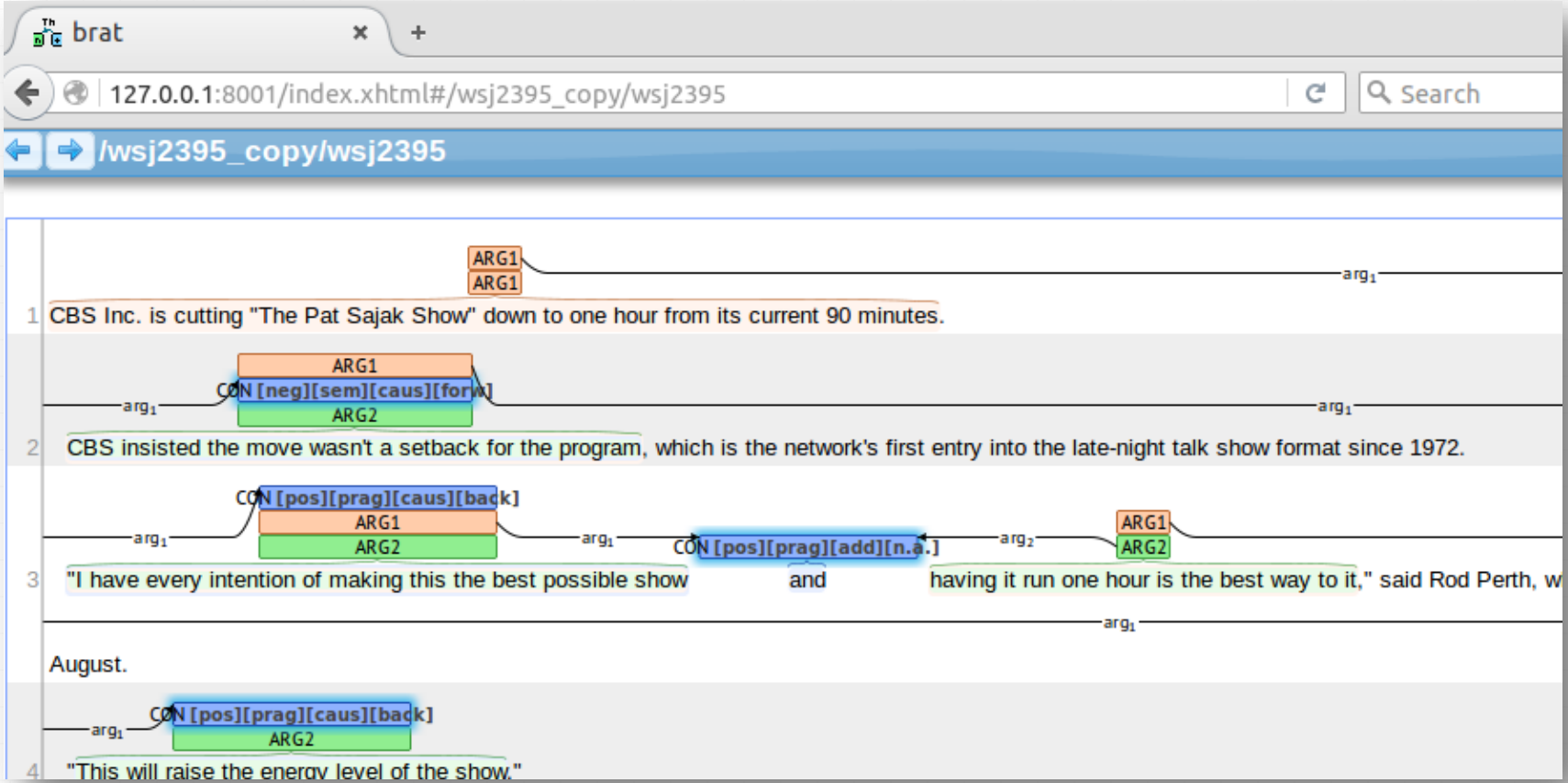
WebAnno



Tools for general aims

server and standalone, web based GUI

brat



Tools for general aims

server and standalone, web based GUI

brat

http://brat.nlplab.org/

ADVANTAGES:

- flexibility defining annotation schema
- some flexibility to adapt visualization
- good and user-friendly user interface
- server and stand-alone
- multilayer?

DISADVANTAGES:

- less control regarding tokenisation, sentence splitting
- not so good as WebAnno regarding managing projects and users
- output format a bit weird

Tools for general aims

standalone

MMAX

http://mmax2.sourceforge.net/

ADVANTAGES:

- flexibility to define own annotation schema
- flexibility to define visualization
- one can keep the original sentence splitting and tokenization of the text
- everything is XML (annotation, text, schema, visualization)
- multilayer annotation

Tools for general aims

standalone

MMAX

http://mmax2.sourceforge.net/

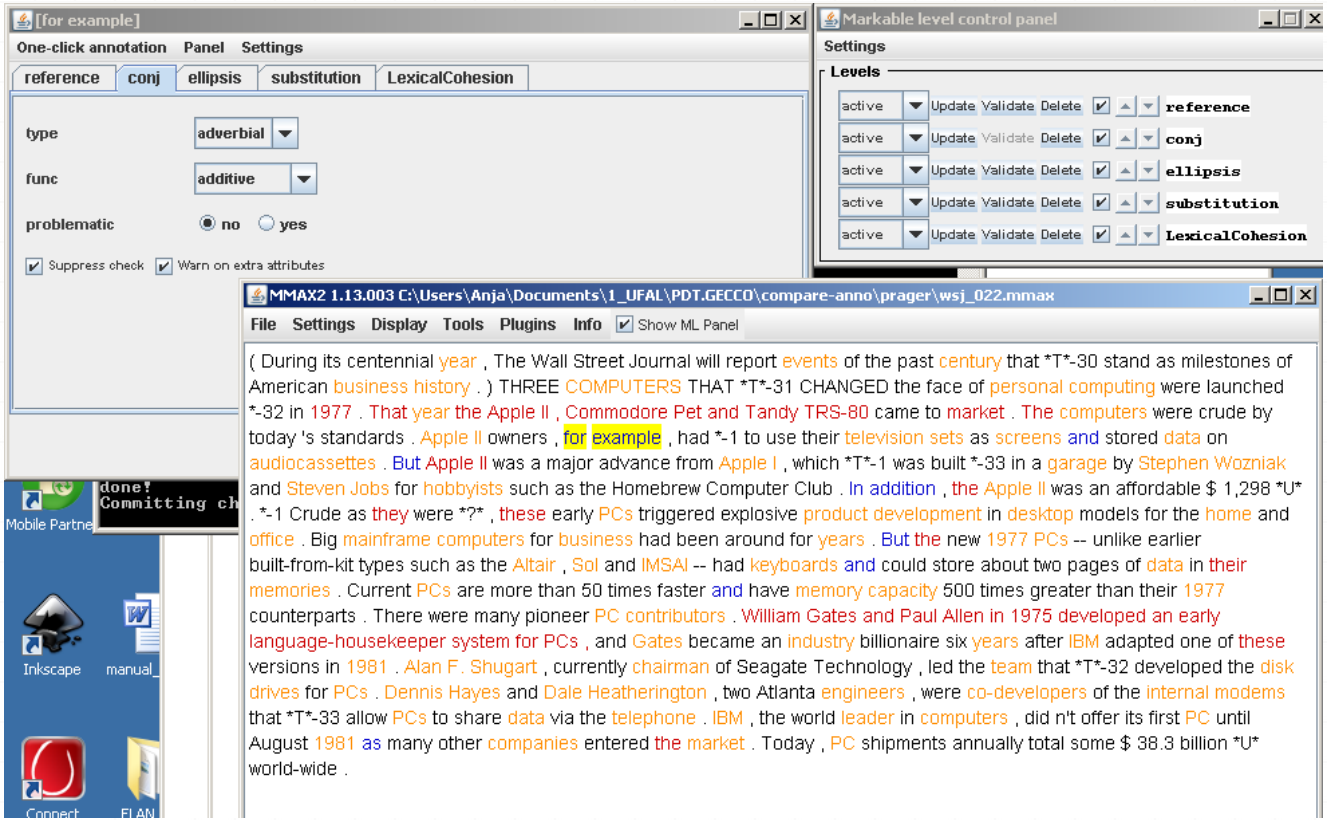
DISADVANTAGES:

- sometimes difficulties to set up the environment
- only one format to import texts and export annotation
- GUI not very user-friendly (lot of clicks and different behaviors depending on the context, frustrating)
- some little details that produce errors (names of folders/files)
- complex data structures exposed to the user (problems if users change the location of the folders where any file is stored)
- layers are unconnected, do not make the structure; does not display any relational information, but only raw text
- not developed any more

Tools for general aims

standalone

MMAX



Tools for general aims

standalone

MMAX

0 GECCo – Lexical cohesion chains

Tools for general aims

standalone

Tree Editor TrEd

http://ufal.mff.cuni.cz/tred/

ADVANTAGES:

- fully customizable and programmable graphical editor and viewer for editing trees
- flexibility defining annotation schema, extensions
- multilayer, interconnected layers
- xml-based, applicable to al possible tree analyses
- extra-powerful search engine

Tools for general aims

standalone

Tree Editor TrEd

http://ufal.mff.cuni.cz/tred/

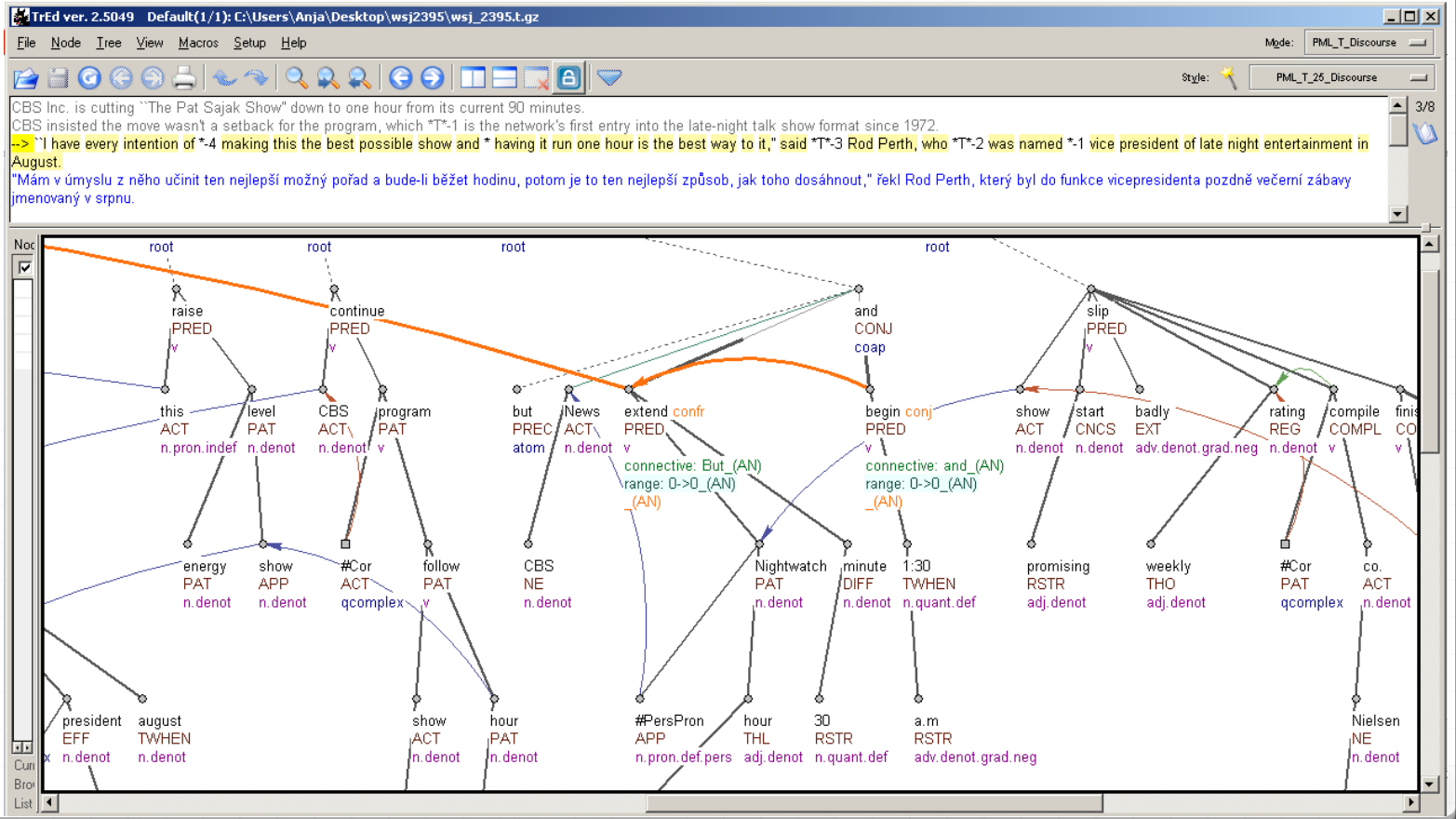
DISADVANTAGES:

- not developed any more
- may be complicated to learn when you annotate things that are not directly connected to tree structure

Tools for general aims

standalone

Tree Editor TrEd



Tools for spoken data

standalone

http://www.exmaralda.org/en

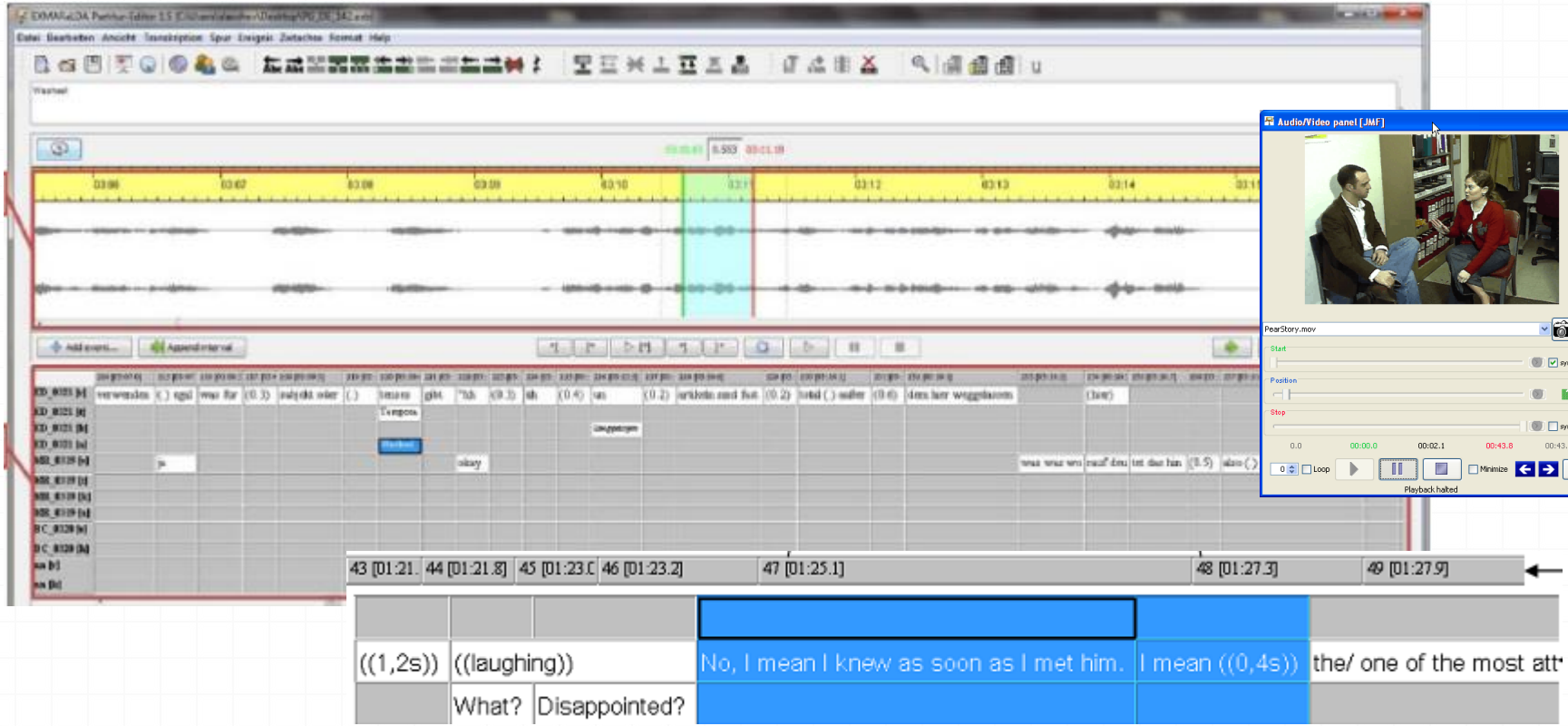
EXMARaLDA

Tool to annotate video-audio files with multiple layers of annotation

ADVANTAGES:

- flexibility defining annotation schema
- specially good for spoken corpora/transcription
- transcription/annotation tool, corpus manager, query interface
- XML-based, compatible with Praat, ELAN, Transcriber
- extensive documentation and tutorials
- good community, help desk
- under active development
- multilayer annotation

EXMARaLDA – partitur editor



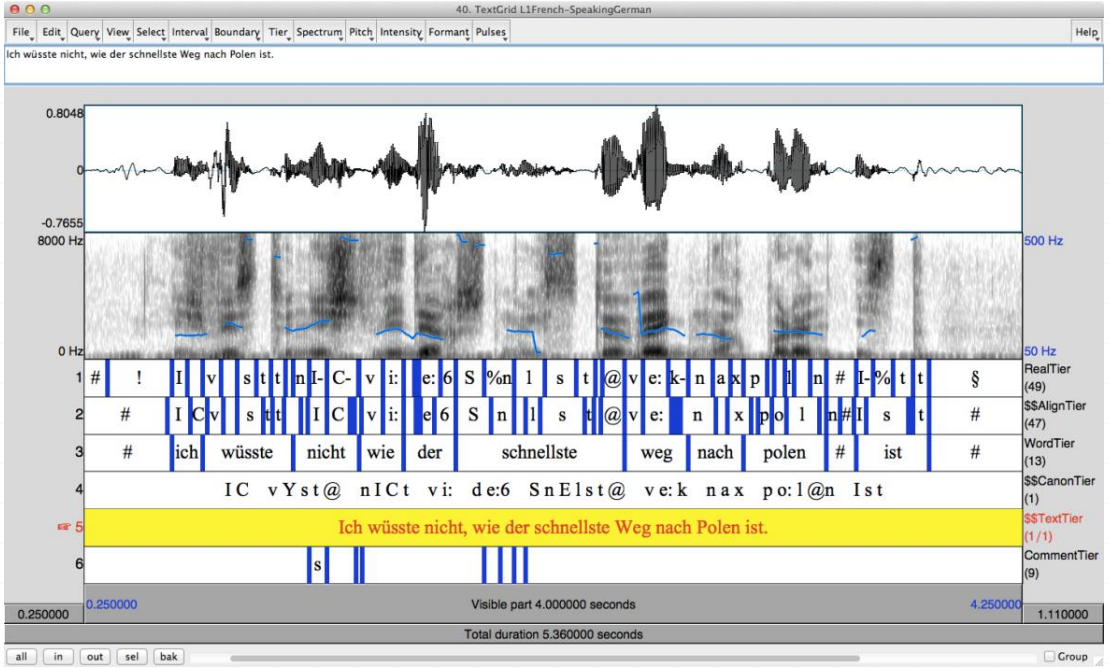
Tools for spoken data

standalone

http://www.fon.hum.uva.nl/praat/

Praat

- Paul Boersma and David Weenink
- Speech analysis and synthesis
- Labelling and segmentation
- Learning algorithms
- Speech manipulation,
- etc.



ELAN

Tool to annotate video-audio files with multiple layers of annotation

- XML
- flexibility
- navigation
- search
- different import/export formats
- documentation
- tool maintained

CBS Inc. is cutting "The Pat Sajak Show" down to one hour from its current 90 minutes. CBS insisted the move wasn't a setback for the program, which is the network's first entry into the late-night talk show format since 1972. "I have every intention of making this the best possible show and having it run one hour is the best way to it," said Rod Perth, who was named vice president of late night entertainment in August. "This will raise the energy level of the show." CBS will continue to program action-adventure shows to follw the Sajak hour. But CBS News will extend its four-hour "Nightwatch" by 30 minutes and begin at 1:30 a.m. The show, despite a promising start, has slipped badly in the weekly ratings as compiled by A.C. Nielsen Co., finishing far below "Tonight" on NBC, a unit of General Electric Co., and "Nightline" on ABC-TV, a unit of Capital Cities/ABC Inc. Further fractioning the late-night audience is the addition of the "Arsenio Hall Show," syndicated by Paramount Communications Inc.

wsj2395

LAB

PDTB tool
RSTweb tool
brat tool

Installation packages and instructions

Install PDTB, RST-web and Brat according to the following instructions (for Linux):

1. Download the archived package from https://ufal.mff.cuni.cz/%7Enedoluzko/tools.tar.gz

2. In your computers, create TOOLS directory and unpack the content of the attached package to it. You will have three directories: PDTB, RST-web and brat.

3. **PDTB**
a) Make sure that Java is installed.
b) Run start.sh to make sure the tool works.

4. **RST-web**
a) Make sure Python 2.X is installed (preferably 2.6 or newer)
b) The Python package cherrypy must be installed if it isn't already (e.g. using pip install cherrypy from the command line)
c) Run start.sh to make sure the tool works.
d) Open rstWeb in your browser at: http://127.0.0.1:8080/ (I use Firefox)

5. **Brat**
a) Make sure Python 2.X is installed (preferably 2.6 or newer)
b) Run start.sh to make sure the tool works.
c) Open brat in your browser at: http://127.0.0.1:8001/ (I use Firefox)
d) To log in, use username: anot, password: anot
In case a red error message in browser arises, ignore it, it seems to have no effect. However, if the tool still doesn't work, run ./install.sh -u in terminal. You will be asked to enter username and password. Use anot, anot, or any other but remember it.

Generally, for all tools, if any errors or problems arise, don't hesitate to describe them to me (nedoluzko@ufal.mff.cuni.cz), and we will try to solve them together.

Technical support

- (right) Ctrl+f : switches to Windows
- pip install **--user** cherrypy
- ./start.sh (in terminal after clicking on start.sh)
- for permissions:

/tools/PDTB# chmod +x start.sh
/tools/PDTB# ./start.sh

Acknowledgements

Many thanks to my colleagues who helped me to prepare this presentation and demonstration, especially to Bonnie Webber, Ágnes Abuczki, José Manuel Martínez Martínez and my husband Dmitry Lukin.