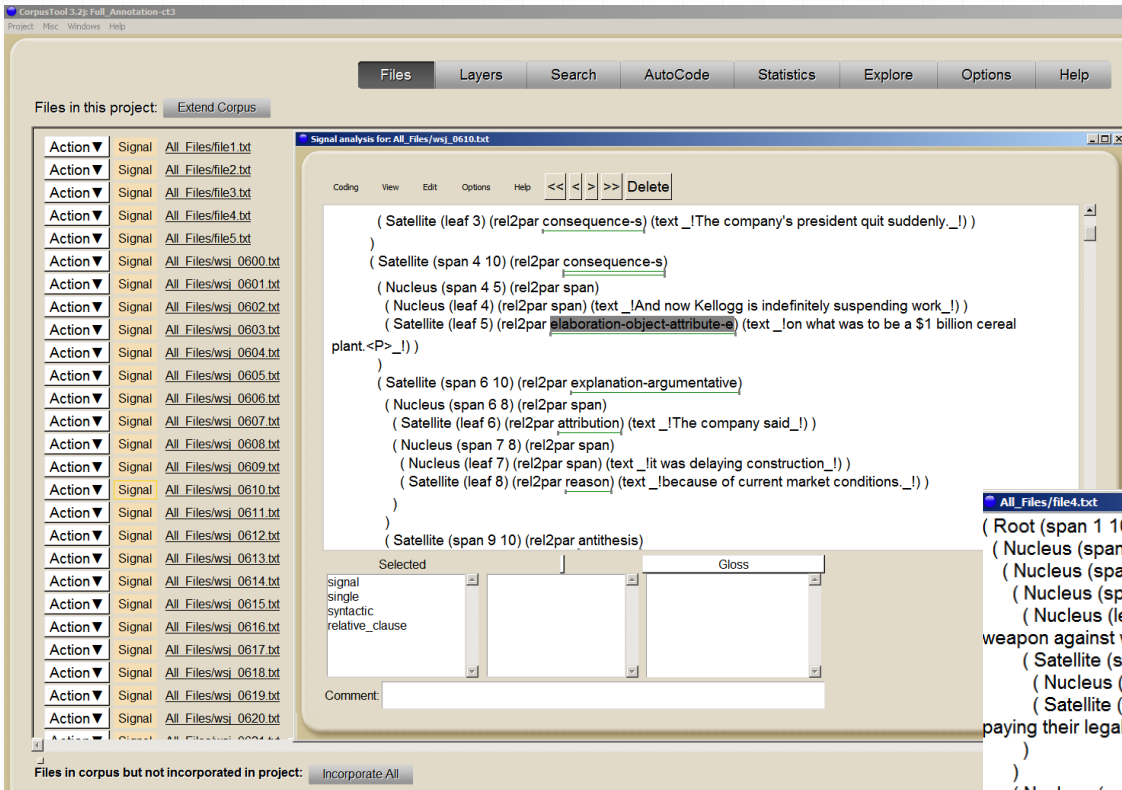


Discourse Annotation Tools overview

Anja Nedoluzhko

Discourse Annotation Tools

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



UAM Corpus Tool3

```

All_Files/file4.txt
( Root (span 1 104)
( Nucleus (span 1 89) (rel2par span)
( Nucleus (span 1 6) (rel2par span)
( Nucleus (span 1 3) (rel2par Cause-Result)
( Nucleus (leaf 1) (rel2par span) (text _!The government is sharpening its newest
weapon against white-collar defendants _!))
( Satellite (span 2 3) (rel2par elaboration-general-specific-e)
( Nucleus (leaf 2) (rel2par span) (text _!the power_!) )
( Satellite (leaf 3) (rel2par elaboration-object-attribute-e) (text _!to prevent them from
paying their legal bills_!))
)
)
)
( Nucleus (span 4 6) (rel2par Cause-Result)
( Satellite (leaf 4) (rel2par attribution) (text _!And defense lawyers are warning_!) )
( Nucleus (span 5 6) (rel2par span)
( Nucleus (leaf 5) (rel2par span) (text _!that they won't stick around_!) )
( Satellite (leaf 6) (rel2par circumstance) (text _!if they don't get paid_!))
)
)
)
( Satellite (span 7 89) (rel2par example)

```

RST Signalling Corpus (Das – Taboada, 2015)

Tools for general aims

standalone

Glozz

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

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

The screenshot shows the Glozz software interface. The main window displays a dialogue transcript with several lines of text, each with a speaker identifier and a timestamp. The text is annotated with colored boxes and lines. A red circle highlights a specific annotation, labeled 'CDU'. The interface includes a toolbar at the top with various icons, a 'Color mode' dropdown set to 'STYLESHEET', and a sidebar on the right with panels for 'Units', 'Relations', and 'Schemas'. The 'Units' panel lists various units like 'Turn', 'Segment', 'Accept', etc. The 'Relations' panel lists relations like 'Anaphora', 'Question-answer', 'Result', etc. The 'Schemas' panel lists schemas like 'Bargaining_block', 'Complex_discourse', etc. Below these panels is a table with columns for 'Sort/Type', 'Sort/Date', 'Show sel.', and 'Visible'. The table contains several rows of data, including 'u_Turn(1,37) ID=1', 'u_Turn(38,56) ID=3', 'u_Turn(57,71) ID=5', 'u_Turn(72,98) ID=7', and 'u_Turn(99,132) ID=9'. A 'Command:' field is at the bottom of the sidebar.

Glozz - 1.1.0-beta - Logged as kowey

Color mode: STYLESHEET

CDU

192 : amycharl : anyone want a sheep

193 : amycharl : ?

194 : IG : sry

195 : sabercat : for what?

196 : amycharl : wheat preferably

197 : sabercat : dont have that :D

198 : amycharl : anything else is fine

Units: Turn, Segment, Accept, Refusal, Strategic_commen, Other, Offer, Countainoffae

Relations: Anaphora, Question-answer, Result, Comment, Continuation, Conditional, Explanation, Elshoration

Schemas: Bargaining_block, Complex_discourse, Several_resources

Sort/Type	Sort/Date	Show sel.	Visible
u_Turn(1,37)	ID=1		
u_Turn(38,56)	ID=3		
u_Turn(57,71)	ID=5		
u_Turn(72,98)	ID=7		
u_Turn(99,132)	ID=9		

Command:

(ideal) PLAN for today

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

Tools – classification criteria

- 0 general

- 0 spoken

- 0 written

- 0 advantages

- 0 disadvantages

- 0 standalone

- 0 server

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

ADVANTAGES:

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

WebAnno

<https://webanno.github.io/webanno/>

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:

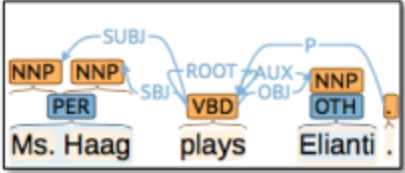
- managing projects with multiple annotators
- curation/revision tasks
- interannotator agreement
- monitoring progress of the project

DISADVANTAGES:

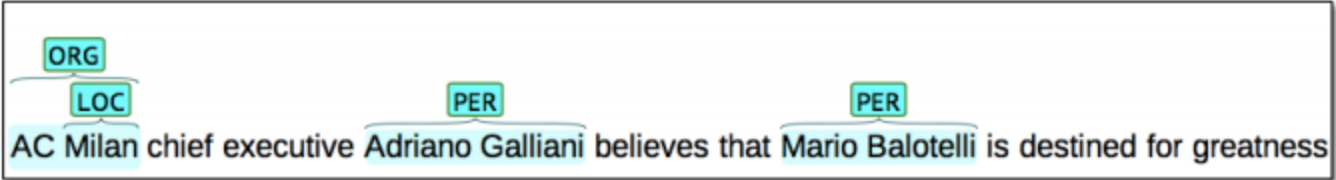
not possible or not easy to annotate some types of information: for example, constituents

WebAnno

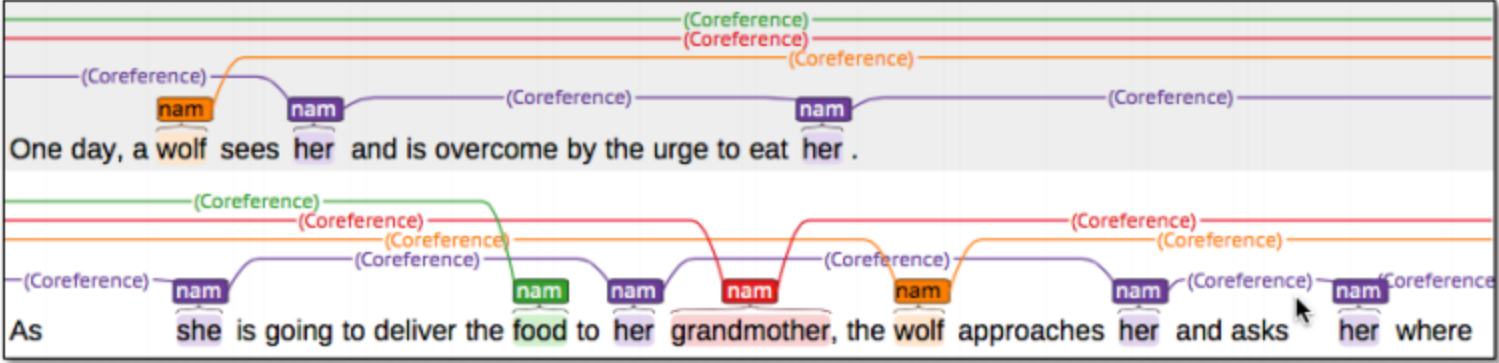
Part-of-Speech & syntactic dependencies



Named entities

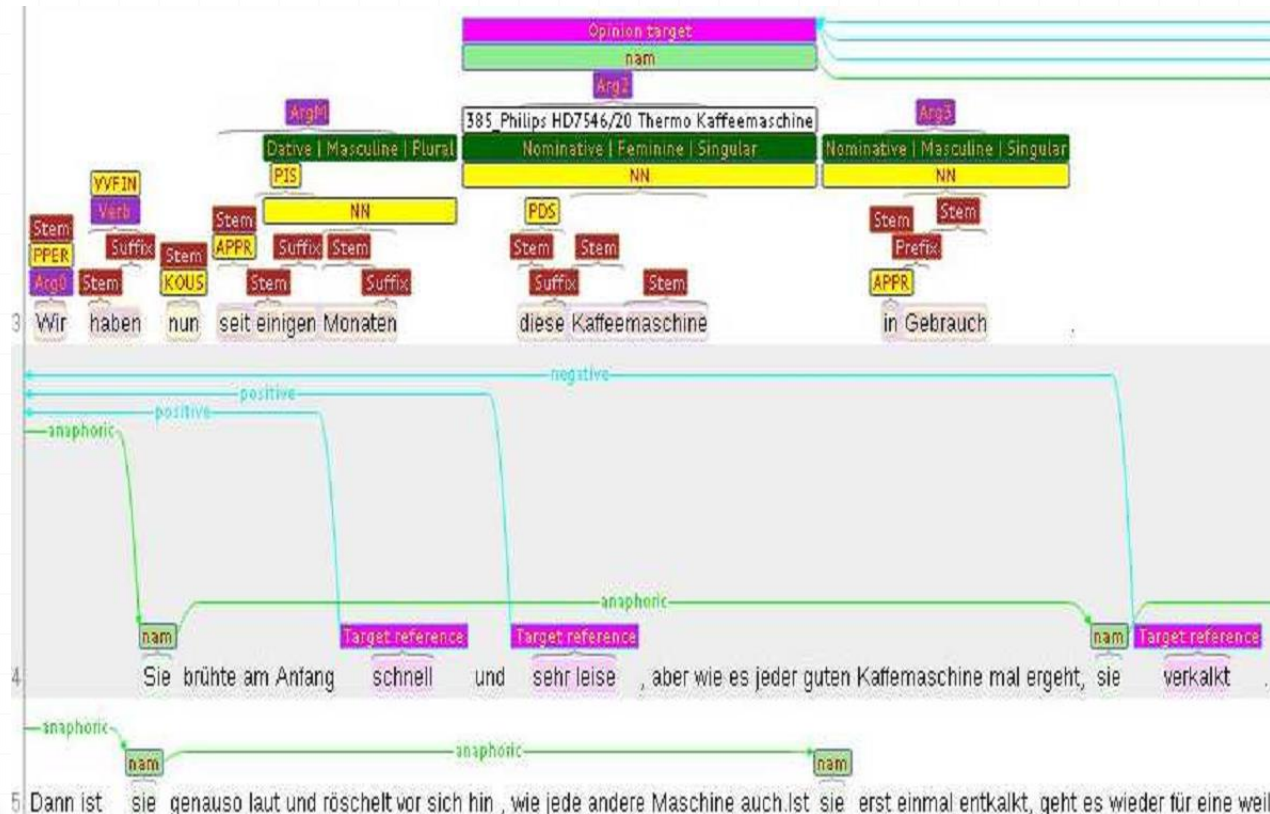


Co-reference



Annotation with multiple layers

WebAnno



Tools for general aims

server and standalone, web based GUI

brat

The screenshot shows the brat web interface in a browser window. The address bar displays the URL `127.0.0.1:8001/index.xhtml#/wsj2395_copy/ws2395`. The main content area displays four sentences from a news article, each with semantic annotations. The annotations include relations (CON) and arguments (ARG1, ARG2) with associated semantic classes (neg, sem, caus, forw, pos, prag, add, n.a.).

1 CBS Inc. is cutting "The Pat Sajak Show" down to one hour from its current 90 minutes.

2 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.

3 "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, w

August.

4 "This will raise the energy level of the show."

Tools for general aims

server and standalone, web based GUI

ADVANTAGES:

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

brat

<http://brat.nlplab.org/>

DISADVANTAGES:

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

Tools for general aims

standalone

MMAX

<http://mmax2.sourceforge.net/>

ADVANTAGES:

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

Tools for general aims

standalone

MMAX

<http://mmax2.sourceforge.net/>

DISADVANTAGES:

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

Tools for general aims

standalone

MMA

The screenshot displays the MMA (Markable MMA) software interface. It features a main window with a menu bar (File, Settings, Display, Tools, Plugins, Info) and a toolbar. The main text area contains a paragraph of text with various words highlighted in different colors (yellow, orange, red, green) to represent different annotation levels. A settings panel is open on the right, showing a list of levels (reference, conj, ellipsis, substitution, LexicalCohesion) with their respective status (active) and actions (Update, Validate, Delete). A smaller settings window is also visible in the top left, showing options for 'type' (adverbial), 'func' (additive), and 'problematic' (no/yes).

[[for example]]

One-click annotation Panel Settings

reference conj ellipsis substitution LexicalCohesion

type adverbial

func additive

problematic no yes

Suppress check Warn on extra attributes

Markable level control panel

Settings

Levels

active	Update	Validate	Delete	<input checked="" type="checkbox"/>	reference
active	Update	Validate	Delete	<input checked="" type="checkbox"/>	conj
active	Update	Validate	Delete	<input checked="" type="checkbox"/>	ellipsis
active	Update	Validate	Delete	<input checked="" type="checkbox"/>	substitution
active	Update	Validate	Delete	<input checked="" type="checkbox"/>	LexicalCohesion

MMA2 1.13.003 C:\Users\Anja\Documents\1_UFAL\ PDT.GECCO\compare-anno\prager\wsj_022.mmax

File Settings Display Tools Plugins Info Show ML Panel

(During its centennial year , The Wall Street Journal will report events of the past century that *T*-30 stand as milestones of American business history .) THREE COMPUTERS THAT *T*-31 CHANGED the face of personal computing were launched *-32 in 1977 . That year the Apple II , Commodore Pet and Tandy TRS-80 came to market . The computers were crude by today 's standards . Apple II owners , for example , had *-1 to use their television sets as screens and stored data on audiocassettes . But Apple II was a major advance from Apple I , which *T*-1 was built *-33 in a garage by Stephen Wozniak and Steven Jobs for hobbyists such as the Homebrew Computer Club . In addition , the Apple II was an affordable \$ 1,298 *U* *-1 Crude as they were ** , these early PCs triggered explosive product development in desktop models for the home and office . Big mainframe computers for business had been around for years . But the new 1977 PCs -- unlike earlier built-from-kit types such as the Altair , Sol and IMSAI -- had keyboards and could store about two pages of data in their memories . Current PCs are more than 50 times faster and have memory capacity 500 times greater than their 1977 counterparts . There were many pioneer PC contributors . William Gates and Paul Allen in 1975 developed an early language-housekeeper system for PCs , and Gates became an industry billionaire six years after IBM adapted one of these versions in 1981 . Alan F. Shugart , currently chairman of Seagate Technology , led the team that *T*-32 developed the disk drives for PCs . Dennis Hayes and Dale Heatherington , two Atlanta engineers , were co-developers of the internal modems that *T*-33 allow PCs to share data via the telephone . IBM , the world leader in computers , did n't offer its first PC until August 1981 as many other companies entered the market . Today , PC shipments annually total some \$ 38.3 billion *U* world-wide .

done! Committing ch

Mobile Partner

Inkscape manual

Connect FL AN

Tools for general aims

standalone

MMAX

0 GECCo – Lexical cohesion chains

If this is special paper such as letterhead, load it with the side to be printed down and the top edge toward the front of the tray.

Note If you are manually duplexing, see "Printing on both sides (duplexing manually)" on page 41 for loading instructions.

5 Make sure the stack of paper is flat in the tray at all four corners, and keep it below the height tabs on the paper length guide in the rear of the tray.

Push down on the metal paper lift plate to lock it into place.

6 Slide the tray back into the printer.

If you set the rear of the tray to one of the longer sizes, the back of the tray will protrude from the back of the printer.

Printing a job This section provides basic printing instructions.

When making changes to printing settings, there is a hierarchy to how changes are prioritized.

(Note that the names of commands and dialog boxes might vary depending on your program.)

"Page Setup" dialog box.

This dialog box opens when you click Page Setup or a similar command on the File menu.

This dialog box is part of the program in which you are working.

Settings changed here override settings changed anywhere else.

"Print" dialog box.

This dialog box opens when you click Print, Print Setup, or a similar command on the File menu.

It is also part of the program, but it has a lower priority than the Page Setup dialog box.

Settings changed in the Print dialog box do not override settings changed in the Page Setup dialog box.

repetition

synonym

holonym

meronym

DEMO

Tools for general aims

standalone

Tree Editor TrEd

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

ADVANTAGES:

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

Tools for general aims

standalone

Tree Editor TrEd

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

DISADVANTAGES:

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

DEMO

Tools for general aims

standalone

Tree Editor TrEd

The screenshot displays the TrEd software interface. The title bar reads "TrEd ver. 2.5049 Default(1/1): C:\Users\Anja\Desktop\wsj2395\wsj_2395.Lgz". The menu bar includes "File", "Node", "Tree", "View", "Macros", "Setup", and "Help". The status bar shows "Mde: PML_T_Discourse" and "Style: PML_T_25_Discourse".

The main text area contains the following text:
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 "T*-1 is the network's first entry into the late-night talk show format since 1972.
-> "I have every intention of *-4 making this the best possible show and *- having it run one hour is the best way to it," said "T*-3 Rod Perth, who *T*-2 was named *-1 vice president of late night entertainment in August.
"Mám v úmyslu z něho učinit ten nejlepší možný pořad a bude-li běžet hodinu, potom je to ten nejlepší způsob, jak toho dosáhnout," řekl Rod Perth, který byl do funkce vicepresidenta pozdně večerní zábavy jmenovaný v srpnu.

The main window displays a complex tree diagram with nodes and edges. The root node is labeled "root". The diagram shows a hierarchical structure of nodes representing the sentence. The nodes are labeled with words and their grammatical categories, such as "raise PRED", "this ACT", "level PAT", "CBS ACT", "program PAT", "and CONJ", "slip PRED", "show ACT", "start CNCS", "badly EXT", "rating REG", "compile COMPL", "finis CO", "energy PAT", "show APP", "#Cor ACT", "follow PAT", "but PREC", "News ACT", "extend PRED", "connective: But_(AN)", "connective: and_(AN)", "begin conj", "show ACT", "start CNCS", "badly EXT", "rating REG", "compile COMPL", "finis CO", "CBS NE", "Nightwatch PAT", "minute DIFF", "1:30 TWHEN", "promising RSTR", "weekly THO", "#Cor PAT", "co. ACT", "president EFF", "august TWHEN", "show ACT", "hour PAT", "#PersPron APP", "hour THL", "30 RSTR", "a.m RSTR", "Nielsen NE".

The diagram is a complex tree structure representing the sentence. The root node is labeled "root". The tree branches out into several main paths. The leftmost path starts with "raise PRED" (v) leading to "this ACT" (n.pron.indef) and "level PAT" (n.denot). The middle path starts with "continue PRED" (v) leading to "CBS ACT" (n.denot) and "program PAT" (v). The rightmost path starts with "and CONJ" (coap) leading to "begin conj" (v) and "slip PRED" (v). The "begin conj" node further branches into "show ACT" (n.denot), "start CNCS" (n.denot), "badly EXT" (adv.denot.grad.neg), "rating REG" (n.denot), "compile COMPL" (v), and "finis CO" (v). The "show ACT" node branches into "energy PAT" (n.denot) and "show APP" (n.denot). The "start CNCS" node branches into "CBS NE" (n.denot) and "Nightwatch PAT" (n.denot). The "badly EXT" node branches into "minute DIFF" (n.denot) and "1:30 TWHEN" (n.quant.def). The "rating REG" node branches into "promising RSTR" (adj.denot) and "weekly THO" (adj.denot). The "compile COMPL" node branches into "#Cor PAT" (qcomplex) and "co. ACT" (n.denot). The "finis CO" node branches into "president EFF" (n.denot) and "august TWHEN" (n.denot). The "energy PAT" node branches into "show ACT" (n.denot) and "hour PAT" (n.denot). The "show APP" node branches into "hour PAT" (n.denot) and "30 RSTR" (n.quant.def). The "Nightwatch PAT" node branches into "#PersPron APP" (n.pron.def.pers) and "hour THL" (adj.denot). The "1:30 TWHEN" node branches into "a.m RSTR" (adv.denot.grad.neg) and "Nielsen NE" (n.denot). The "#Cor PAT" node branches into "show ACT" (n.denot) and "hour PAT" (n.denot). The "co. ACT" node branches into "show ACT" (n.denot) and "hour PAT" (n.denot). The "show ACT" node branches into "hour PAT" (n.denot) and "30 RSTR" (n.quant.def). The "hour PAT" node branches into "a.m RSTR" (adv.denot.grad.neg) and "Nielsen NE" (n.denot). The "30 RSTR" node branches into "a.m RSTR" (adv.denot.grad.neg) and "Nielsen NE" (n.denot). The "a.m RSTR" node branches into "Nielsen NE" (n.denot). The "Nielsen NE" node branches into "Nielsen NE" (n.denot).

Tools for spoken data

<http://www.exmaralda.org/en>

standalone

EXMARaLDA

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

ADVANTAGES:

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

EXMARaLDA - partitur editor

The screenshot shows the EXMARaLDA Partitur-Editor 1.5 interface. At the top, there's a menu bar with options like 'Datei', 'Bearbeiten', 'Ansicht', 'Transkription', 'Spur', 'Eingabe', 'Zutasten', 'Format', and 'Hilfe'. Below the menu is a toolbar with various icons for editing and playback. The main area features a timeline with a yellow background and a green highlight at 02:11. Below the timeline are two waveform tracks. At the bottom, there's a transcript table with columns for time and text.

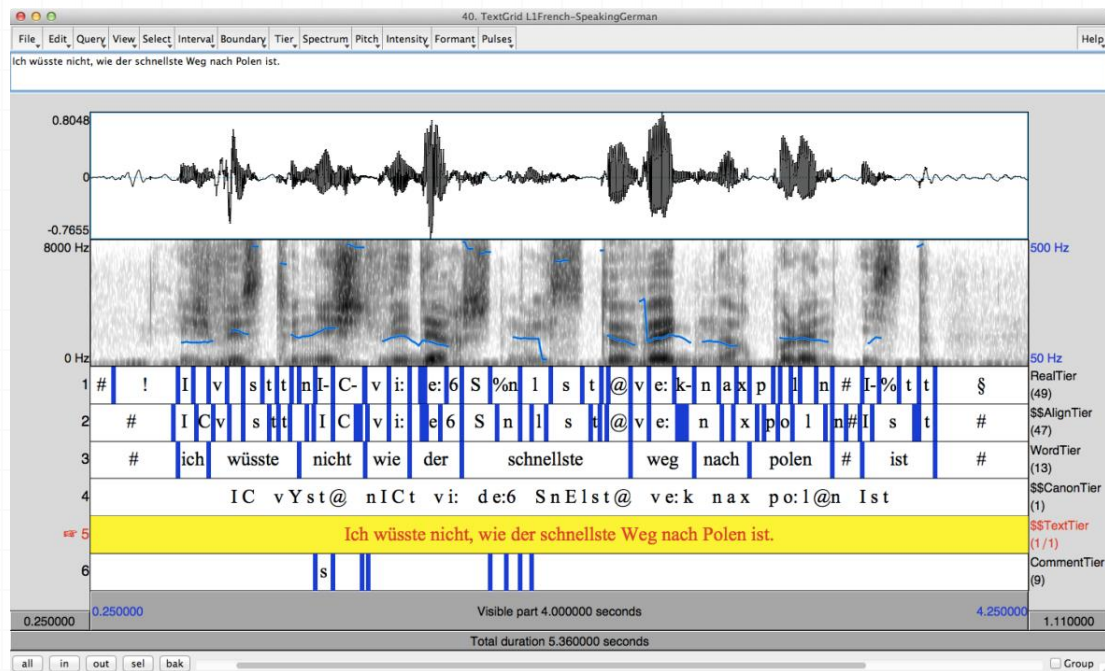
43 [01:21.1]	44 [01:21.8]	45 [01:23.1]	46 [01:23.2]	47 [01:25.1]	48 [01:27.3]	49 [01:27.9]
--------------	--------------	--------------	--------------	--------------	--------------	--------------

The Audio/Video panel [JMF] is a floating window showing a video player. The video displays two people sitting and talking. Below the video are playback controls including 'Start', 'Position', 'Stop', and a progress bar. The progress bar shows a current time of 00:00.0 and a total duration of 00:43.8 sec. There are also buttons for 'Loop', 'Minimize', and 'Playback-halted'.

((1,2s))	((laughing))	No, I mean I knew as soon as I met him.	I mean ((0,4s))	the/	one of the most att
	What?	Disappointed?			

Praat

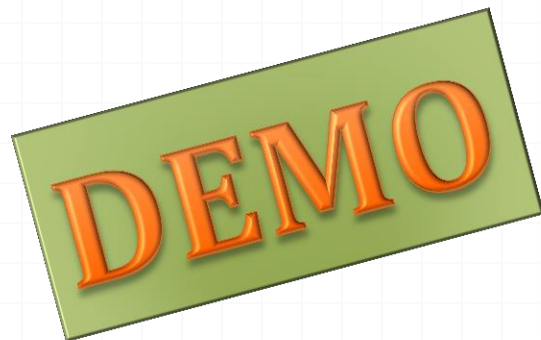
- 0 Paul Boersma and David Weenink
- 0 Speech analysis and synthesis
- 0 Labelling and segmentation
- 0 Learning algorithms
- 0 Speech manipulation,
- 0 etc.



ELAN

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

- 0 XML
- 0 flexibility
- 0 navigation
- 0 search
- 0 different import/export formats
- 0 documentation
- 0 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 follow 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.

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

- 0 (right) Ctrl+f : switches to Windows
- 0 pip install --**user** cherrypy
- 0 ./start.sh (in terminal after clicking on start.sh)
- 0 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.