

Arne Neumann, arne.neumann@uni-potsdam.de

Amir Zeldes, amir.zeldes@rz.hu-berlin.de

Florian Zipser, f.zipser@gmx.de

ANNIS 3

Challenges and Innovations for Corpora in the SFB

D1 in Phase 3

- More support for multimodal corpora
- Moving beyond corpora with single token layers
- More flexible visualization capabilities
- Focus on document-sized contexts
- Keeping everything compatible
- Sustainability

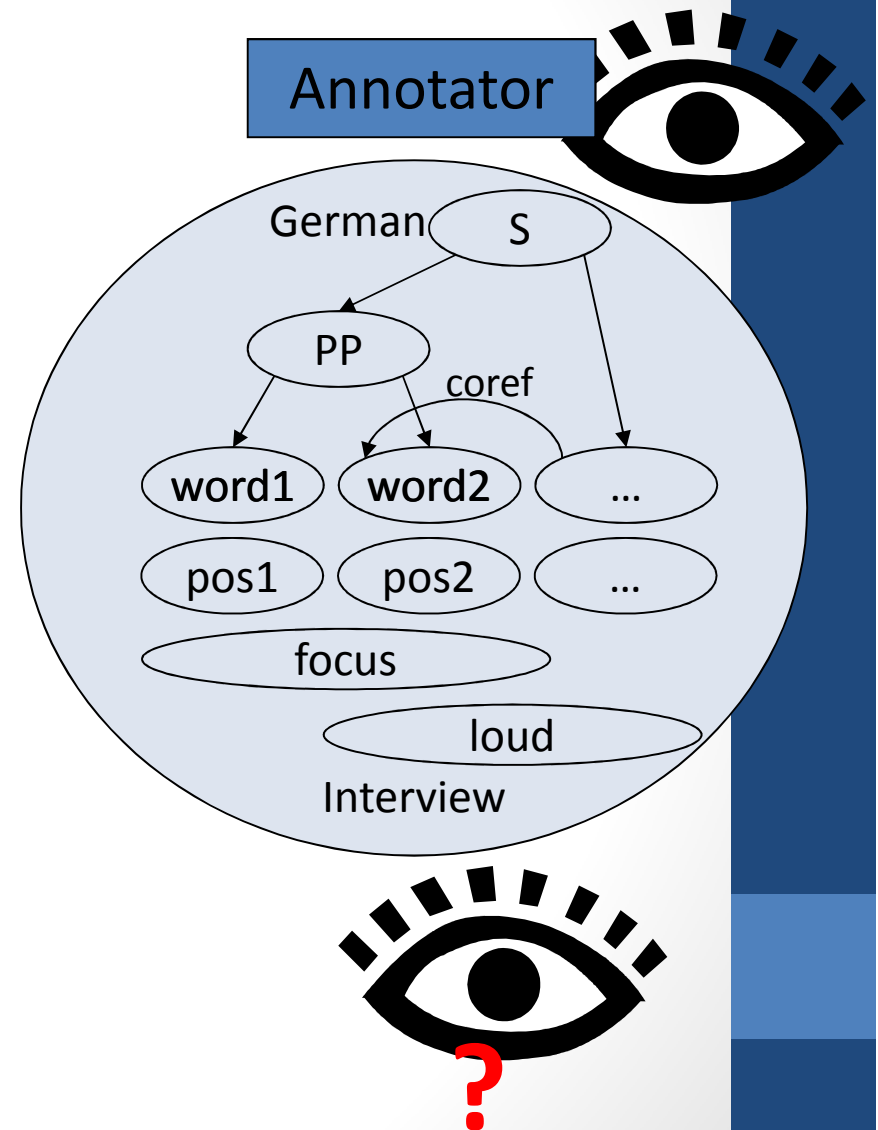
ANNIS



- A standardized solution for all SFB corpora
 - Browser based (nothing to install)
 - Open source (collaborative development)
 - Customizable search and visualization
- Multilayer annotation "extreme"
 - Combine IS annotation, syntax trees (constituents / dependencies), morphology, coreference, RST, multimodal data...

Everything is a graph

- Reduction of all annotations to typed and labeled nodes and edges
- Independent layers (stand-off)
 - Retroactively insert / update / replace layers
 - Alternative interpretations (multiple pos-tags, parses, error analyses...)
 - PAULA XML (**new:** Version 1.1)



ANNIS2

ANNIS2 Tutorial

Search Form

AnnisQL: `[tok & tok & #1 ->dep [func="OA"] #2 & cat="S" & #3 _ #1 & node & #3 >secedge #4 | correction="correcting" | cat="c"]`

Query Builder: [Show >>](#)

Result: 43

History: [Query History](#)

More Corpora

Name	Texts	Tokens
FalkoEssayL2V2_0	248	131511
ONTONOTES_y1.5_small	4	6450
SMULTRON_Banana	2	3782
TueBa5_no_cyc	2187	770849
agni_I	24	184
h4.tation2.0	2031	11295
pcc-3	3	573
pcc2	2	399
<input checked="" type="checkbox"/> tiger1.dep	1	829
<input type="checkbox"/> tiger2	1971	888578

[Search](#) [Export](#)

Context Left: 0

Context Right: 0

Results per page: 10


[Show Result](#)

Search Result - tok & tok & #1 ->dep[func="OA"] #2 & cat="S" & #3 _ #1 & node & #3 >secedge #4 (0, 0)

Page 1 of 5

Token Annotations Show Citation URL

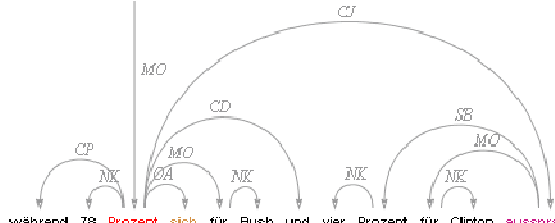
Displaying Results 1 - 10 of 43



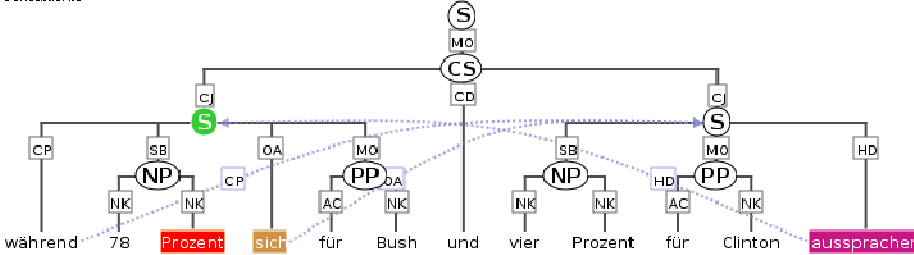
während 78 Prozent sich für Bush und vier Prozent für Clinton aussprechen

während 78 Prozent sich für Bush und vier Prozent für Clinton aussprechen
 KOUS CARD NN PRF APPR NE KON CARD NN APPR NE VVFIN
 -- -- ^.^Neut 3.Acc.Pl -- Acc.Sq.^ -- -- ^.^Neut -- Acc.Sq.^ 3.Pl.Past.Ind

dependencies



constituents



Die Vase auf dem Tisch ist größer als die Vase

animacy (grid)

Select Displayed Annotation Levels

mmax:ref_type	inanim	inanim
mmax:ref_type	inanim	inanim
tok	Die Vase auf dem Tisch ist größer als die Vase	

coreference (discourse)

Die Vase auf dem Tisch ist größer als die Vase auf der Fensterbank . Ich finde , sie sieht nicht so gut aus , weil der Tisch zu klein ist.

New corpora, new challenges

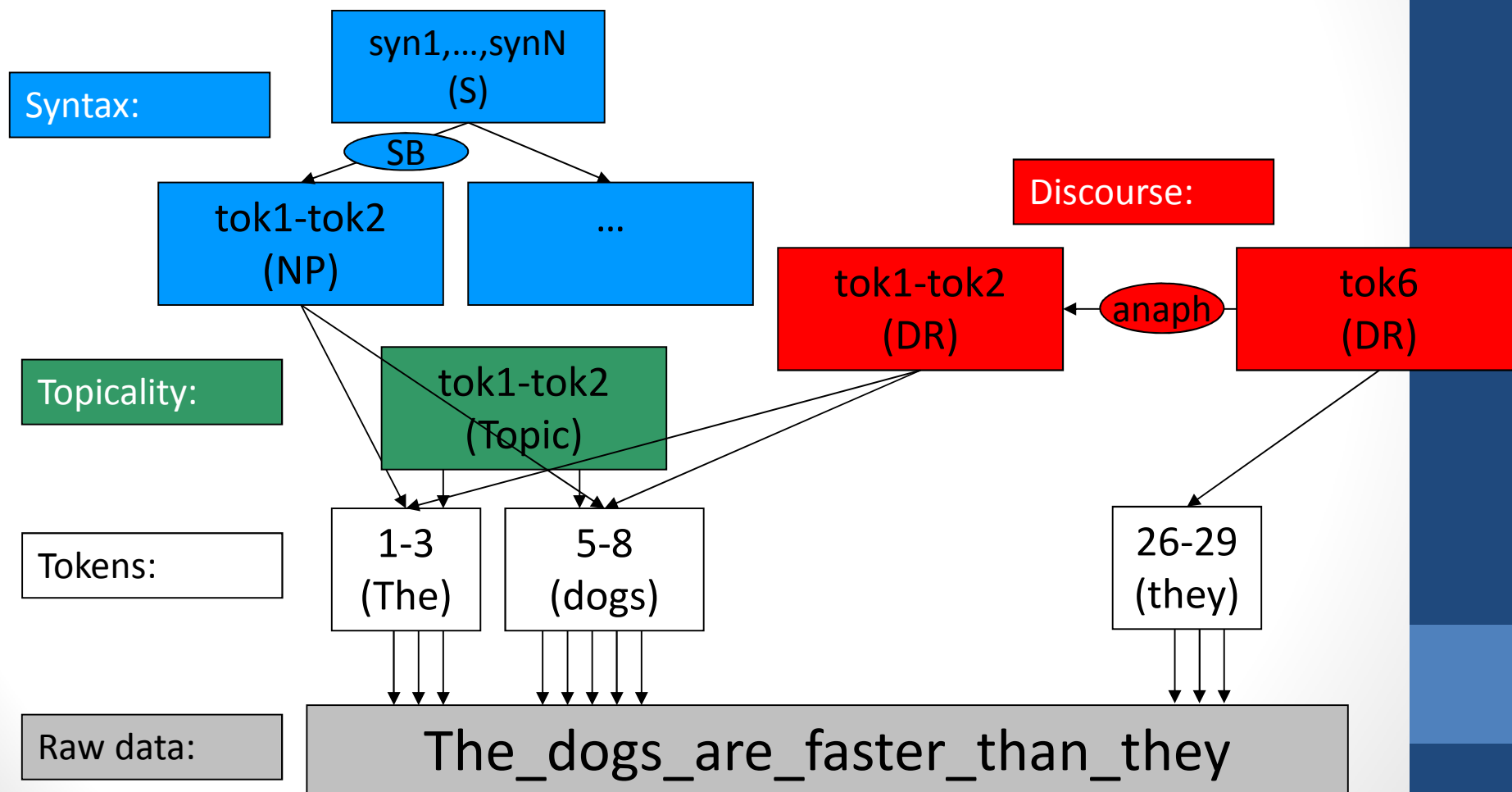
- Spontaneous dialogue data – the Kiezdeutsch corpus **KidKo (B6)**
- Previous multimodal corpora were collected via QUIS, laboratory recordings
- Dialogue ordered, no overlapping speakers

New corpora, new challenges

- Tokenization: segmentation of the corpus text into minimal annotatable units
 - Units are ordered: *<the, dogs, are, faster,...>*
 - No gaps – empty elements *are* tokens
 - Achieved via PAULA XPointer references to text spans

PAULA stand-off XML 1.0 (simplified)

(Dipper, 2005, Dipper & Götze, 2005)

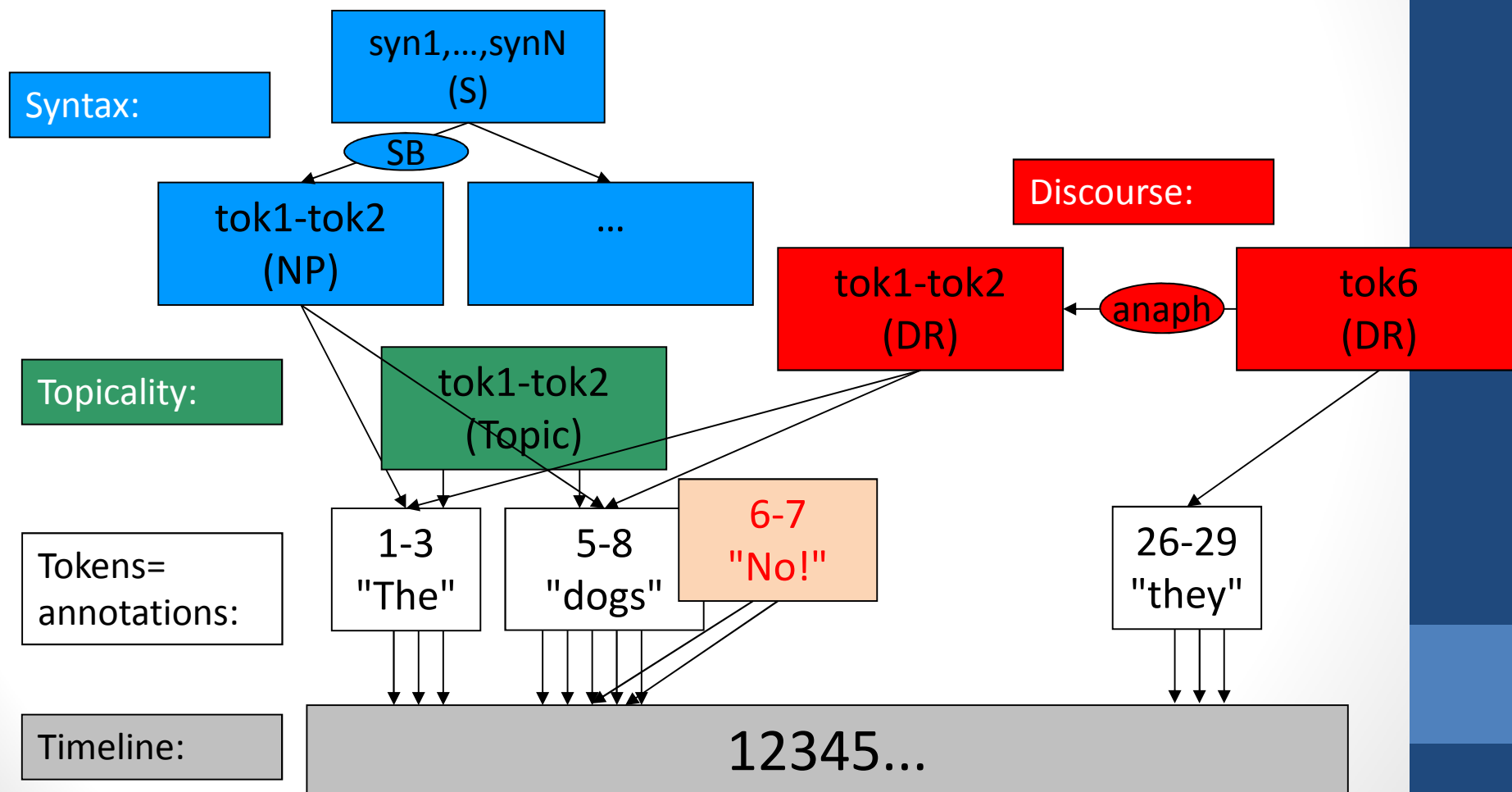


Multiple speakers

- Now we have data from multiple, possibly simultaneous speakers
- Multiple token layers
- How can we adapt the format without disrupting existing structures?

PAULA stand-off XML 1.1 (simplified)

(Zeldes et al. 2013)



Advantages

- Remain within PAULA framework
- Exploit entire power of the graph structure
- Data from each speaker can carry the full load of PAULA annotations available so far (syntax trees, IS...)

Problems

- **We** know that these are transcriptions of speaker data
- **ANNIS** doesn't:
 - What happens when we search for "dog"?
 - How do we find adjacent words?

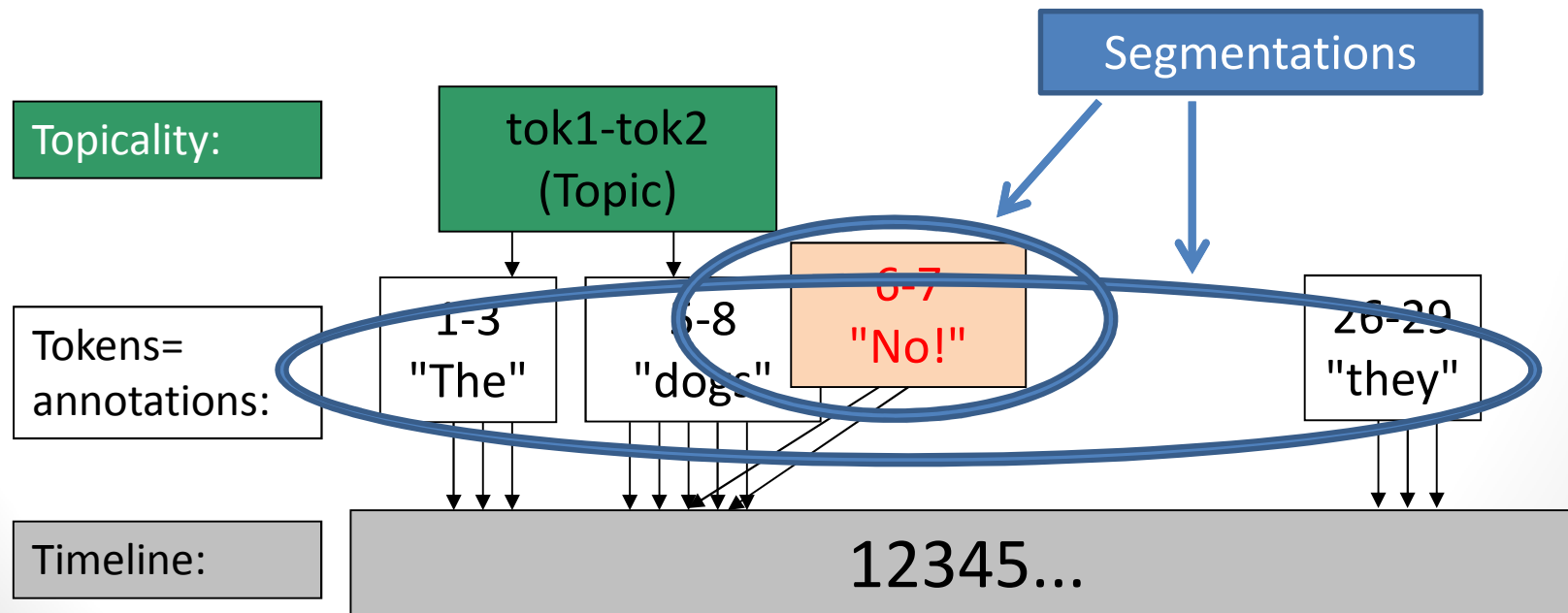
dogs  *are*

- How do we search within n to m words?
- How do we show a context of ± 5 words?

Multiple segmentations

(Krause et al. 2012)

- Tag some layers as "segmentation" layers
 - respond to textual search (someone saying "dog")
 - Used to define context, distance/adjacency between elements



Setting segmentations

Search Form

AnnisQL:

Show Result History

Status: 3 matches in 1 document

Corpus List

Search Options

Left Context: 5

Right Context: 10

Show context in:

- tokens (default)
- tokens (default)
- Vxxx_tr
- Vxxx_v
- Axxx_v
- Exxx_v
- MuH19WT_v**
- Exxx_tr

Results Per Page

Path: b6.MuH19WT_03_POS_NK > Kiezdeutsch2008

MA:NN wir sind doch zu SPÄT gekomm und

MuH19WT_v					
MuH19WT_norm					
MuH19WT_POS					
Vxxx_v	deñ	LO:S			(fremdsprachlich, türkisch)
Vxxx_norm	deñ	los			FOREIGN
Vxxx_POS	ADV	PTKVZ			XYU
Vxxx_tr				noldu	ki
Vxxx_trnorm	Ne	oldu		ki	
Vxxx_trdtwue	was	ist passiert		Partikel	
Vxxx_trdtue				Was ist denn passiert?	
Exxx_v		HÄ			
Exxx_norm		Hä			
Exxx_POS		PTKQU			

audio 2:08

Path: b6.MuH19WT_03_POS_NK > Kiezdeutsch2008

(-) anoNYM (-) auf DEUTSCH reden (-)

MuH19WT_v	(-)	anoNYM	(-)		
MuH19WT_norm	PAUSE	Anonym	PAUSE		
MuH19WT_POS	PAUSE	ADJD	PAUSE		
Exxx_v					(fremdsprachlich, türkisch)
Exxx_norm					FOREIGN
Exxx_POS					XYU
Exxx_tr		salla			
Exxx_trnorm		Salla			

Configurable Visualization

- Previous phase concentrated on search result visualization – syntax trees, dependencies, annotation grids
- In this phase – more focus on **full document visualizations** (close reading)
- Highlighting of interesting features/combinations
- Easier customization of visualizations

Creating ANNIS based HTML

- Concept: let specific corpora configure their own visualizations
- Annotations and values trigger generation of HTML elements
- CSS stylesheets linked to the visualization determine display properties

Some test cases

- A5 & collaborators – Hausa corpora:
 - Hausa news corpus (Deutsche Welle)
 - Umarnin Uwa – an annotated Nollywood film
- B7 & collaborators – Wolof corpora:
 - Parallel Bible corpus (English – Wolof)
 - Translated Wikipedia articles
 - Crawled Web corpus
- D1 – PCC corpus (new visualizations)

A5 Hausa

- Corpora created in cooperation with Ines Fiedler, Katharina Hartmann and Malte Zimmermann
- Bible corpus (from Phase 2) and news texts from Deutsche Welle used to develop a part-of-speech tagset and training corpus
- New corpus of film transcripts (similar to annotated subtitles)

Hausa POS tagging

- Current results:
 - final tagging scheme developed with 32 POS tags
 - so far 93.4% accuracy (baseline 33.4%)
 - training corpus set to grow (currently just ~5000 manually tagged tokens)

Hausa POS tagging

- Possible to search for:
 - focus particles
 - independent personal pronouns (e.g. contrast focus)
 - PACs (a.k.a. TAMs)
 - Negations
 - ...

Hausa news corpus

About ANNIS | Report Bug | Help us to make ANNIS better!

Search Form

AnnisQL: `POS!="N" & POS="FOC" & #1 . #2`

Show Result | History

Status: 5 matches in 4 documents

Corpus List

Visible: All

Name	Texts	Tokens
a5.hausa.news	4	2,017
a5.hausa.umarnin.uwa_V2	47	10,194
abraham.our.father	7	7,705
AP.172.antonius	1	64
apophthegmata.patrum.5	5	700
apophthegmata.patrum.par.	2	416
b6.MuH19WT_03_POS_NP	1	650
b7.wolof.gospels_V2	89	94,132
b7.wolof.web	5	15,335
b7.wolof.wiki.V3	14	12,738

Query Result

Results 1 - 5 of 5

1 Path: a5.hausa.news > babban_taron

coci a Munich Kungiyoyin addini kalilan ne ke iya yin wani abin
N PRP N N N QUANT FOC PROG V V PIND N

2 Path: a5.hausa.news > babban_taron

Yakoubou wannan kusanci na addini
N PDEM N LINK N

3 Path: a5.hausa.news > sabon_bayani

, amma ba wai sun tsara ne tun gabannin lokacin ba
PUNCT CONJ NEG PTC PAC V FOC PRP N N NEG PUNCT

4 Path: a5.hausa.news > soke_tashin

, jaridar Die Tageszeitung ta tsara ne da cewa rufe sararin samaniyar
PUNCT N FM FM PAC V FOC CONJ V V N N

5 Path: a5.hausa.news > tarihin_tsohon

Babangida Musa 'Yar'Adua an haifeshi ne a ranar 16 ga watan
N N N PUNCT PAC V FOC PRP N NUM PRP N

non-post nominal focus particles

... but they allegedly did not deport FOC [them] since ...

Double negation

Film corpus: Umarnin Uwa

- Corpus subdivided into scenes
- Multiple speakers in each scene
- Manual annotation of foreign words and extralinguistic events
- Automatic part-of-speech tagging (with errors!)
- Script like visualization (good for subtitle corpora, turn-based dialogue, theater plays...)

Film corpus: Umarnin Uwa

example queries Tutorial Query Builder Query Result

Base text Token Annotations

1 < 1 / 3 > >| Displaying Results 1 - 10 of 23

1 Path: a5.hausa.umarnin.uwa_V2 > scene_03

cewa za sub a da invitation nan za su ba invitation
V FUT N PRP PRP N PTC FUT PPRS NEG N

☐ annotations (grid)

info	on the phone									
lang					foreign					foreign
scene	scene_03									
speaker	Ibrahim									
tok	cewa	za	sub	a	da	invitation	nan	za	su	

☐ full scene (annotated dialogue)

Ibrahim: ka gane abin da nake so ka fahimta ? Yaya za a yi mutanen nan , tan
yaushe suke cewa za sub a da invitation nan za su ba invitation nan har
yanzu ba su bayar ba ? *if they are not serious* kawai mu dauki abin nan mu
kai wani guri Yanzu ma kwanaki nawa ? Har man maneji yana neman ya
kare mana ? look , abin da nake so ka gane shi in ba su ba ni invitation
din nan yau ba , na hakura kawai . Zan kai wani guri a mun . *that's it* .
To , shi ke nan ina saurarenka .

on the phone

Ummi: Salamu alaikum

Ibrahim: Amin wa alaiku mus salamu

Ummi: Yaya , wai ka zo in ji Hajiya

Ibrahim: To

loan words

code switching

B7 Wolof

- Corpora part-of-speech annotated and translated by Sheikh Bamba Dione (Bergen) in cooperation with B7
- Development of automatic part-of-speech tagging
- Harvesting and annotating a corpus from the Web (resources for Wolof are scarce!)

Wolof Bible chapter view

example queries Tutorial Query Builder Query Result

Base text Token Annotations

Displaying Results 3041 - 3050 of 15325
 Result for query

3041 Path: b7.wolof.gospels_V2 > doc27

daan nañu ko , mu **am** naqaru xol . mu dem
 VVBZ INF PRO \$, PRS VVBZ NVPS NC \$. PRS VVBZ

translation (grid)
 full chapter (text & translation)

3042 Path: b7.wolof.gospels_V2 > doc27

am naqaru xol . mu **dem** nag ca **saraxalekat** yu mag
 VVBZ NVPS NC \$. PRS VVBZ IJ AP NC PREL VVBZ

translation (grid)

trans	then judas , who betrayed him , when he saw that jesus was condemned , felt remorse , and brought back the thirty pieces of silver to the chief priests and elders ,
verse_id	MATT27:3
tok	am naqaru xol . mu dem nag ca saraxalekat yu mag

full chapter (text & translation)

MATT27:1: ci suba teel **saraxalekat** yu mag yépp ak njiiti xeet wa gise , ngir fexee reylu yeesu .

MATT27:2: ñu yeew ko , yóbbu ko , jébbal ko pilaat boroom réew ma .

MATT27:3: bi nga xamee ne yudaa , mi woroon yeesu , gis na ne , daan nañu ko , mu am naqaru xol . mu dem nag ca **saraxalekat** yu mag ya ak njiit ya , delloo leen fanweeri poset yu xalis ,

MATT27:4: ne leen : « am naa bàkkaar ci li ma wor nit ku tooñul , ngir ñu rey ko . » ñu tontu ko : « lu ciy sunu yoon ? loolu yaa ko yég . »

MATT27:5: noonu yudaa sànni xalis ba ca kër yàlla ga , jóge fa , dem , ta **saraxalekat** saying , ' i have sinned in that i betrayed innocent blood . ' but they said , ' what is that to us ? you see to it . '
 baatam , xaru .

MATT27:6: **saraxalekat** yu mag ya nag fab xalis ba , naan : « jaaduul nu def xalis bii ci dencukaayu xalis bi ci kër yàlla gi , ndaxte njègu deret la . »

MATT27:7: ñu daldi diisoo nag , jënd ca xalis ba toolu defarkatu ndaa ya , ngir di fa suul doxandéem ya .

Wolof Wikipedia corpus

Bennoo gu Almaañ Almaañ njëkk xarey
 NC PRON NAME NAME PREP NC

☐ sentences (grid)

English	Unification Of Germany	In the first Napeleonic wars, Germany was approximately composed of 360 states which belonged to the Austria Emperor or to the so called Holy Roman Emperor.
Wolof	Bennoo gu Almaañ	Almaañ njëkk xarey Napoleon yi , dafa séddaliku woon ci lu jeye ñatti téemeer ak juroom benn fukki diiwaan (360 diiwaan) , yépp it ci ron imbraatóor gu Otris lañu nekkoon , walla li ñuy tudde Imbraatóor gu Rom gu sell gi .
tok	Bennoo gu Almaañ	Almaañ njëkk xarey

☐ article (text & translation)

Bennoo gu Almaañ Almaañ njëkk xarey Napoleon yi , dafa séddaliku woon ci lu jeye ñatti téemeer ak juroom benn fukki diiwaan (360 diiwaan) , yépp it ci ron imbraatóor gu Otris lañu nekkoon , walla li ñuy tudde Imbraatóor gu Rom gu sell gi .

it genn nguur rekk a fa nekkoon gu mag te jëmooon na kanam , te ag yewwuteem fés mag mi (Great Fredrick) (1740 - 1786 g) am doole . Bu ko defee waa Almaañ yépp wëlbatu seeni gët jëme ci nguur gii , ngir bëgg ag bennoo . Napoleon ak xarekatam ya xuusoon nañu ci diiwaani Almaañ yi , teg leen loxo . Daldi jawali ca nguurug Brusiya , gi nga xam ne jammaarloo woon na ak ñoom cig njammaar ak fit , waaye Napoleon moom mujj na ko noot , te gaffe ko ci mujj gi , ca xareb Yena ba , atum 1806 g , daldi teg loxo péeyam ba Berlin . Napoleon merloo woon na lool waa Brusiya yi , ci dog gi mu dogoon ay wàll ci seeni suuf , te jébbaloon leen diiwaanu Saksoniya bi . Waaye nag ci geneen wàll , am na lu am njariñ lu mu fa indi , ndaxte wàññi woon na limub diiwaan yu Almaañ yi ba ci fanweer ak

In the first Napeleonic wars, Germany was approximately composed of 360 states which belonged to the Austria Emperor or to the so called Holy Roman Emperor.

Potsdam Commentary Corpus

(Stede 2004)

- corpus of German newspaper commentaries
- annotated on multiple levels

Does Zossen need a new youth club?



Should this building be demolished?



PCC: Size

- 176 German newspaper commentaries from *Märkische Allgemeine Zeitung* (local daily newspaper)
- 44 pro and contra editorials from *Der Tagesspiegel* (a regional daily)
- new complete PCC release incl. IS and all other available annotations coming soon

PCC: Annotations

Four layers of manual annotation

- POS, morphology and syntax (STTS, Schiller et al. 1999 and TIGER scheme, Brants et al. 2002)
- Coreference & entities (PoCoS scheme, Krasavina and Chiarcos 2005, upd. 2012)
- Rhetorical Structure Theory (Mann and Thompson 1988)
- Information structure (Dipper et al. 2007, WIP)

PCC: ProCon 10 subcorpus

Additional annotations

- content zones
- topics
- information structure
- conjunctive relations
- illocutionary status
- argumentative structure

IS: Visualization

Feigenblatt

[Die Jugendlichen in Zossen] [wollen ein Musikcafé] . Das forderten [sie] [bei der ersten Zossener Runde am Dienstagabend] . Dass [die Politiker der Stadt] dafür Verständnis haben , ist loblich . [Mit dem Treffen im Rathaus] [ist somit auch ein Dialog zwischen den Generationen angestoßen] . Dass [die beiden geladenen Jugendlichen] [im Laufe des Abends] [immer weniger zu Wort kamen , war sicher keine böse Absicht , ärgerlich ist es trotzdem] . [Und aberwitzig dazu] . Es entbehrt nicht der Komik , wenn sich [drei Erwachsene - Karola Andrae (Bürgerbündnis/FDP) , Susanne Michler (CDU) und Joachim Zanow (SPD) -] darüber streiten , was Jugendliche wollen und brauchen , ohne auf die Idee zu kommen , sie selbst zu fragen . Und das , obwohl [sie] ihnen gegenüber sitzen . [Die Jugendlichen] [wurden somit zum bloßen Feigenblatt degradiert] . [Nicht über sondern mit ihnen hätten] [die Politiker] [reden sollen] . Damit ist eine große Chance vertan . Vielleicht klappt es [bei der nächsten Runde Anfang 2002 .] [Dann] werden auch mehr Jugendliche eingeladen . [In der Gruppe können] [sie] [sich hoffentlich mehr Gehör verschaffen] . [Vielleicht finden dann auch] [Vertreter von PDS und Gewerbeverein] [ihren Weg ins Rathaus] . [Die] [glänzten diesmal noch mit Abwesenheit] .

- information status: underlined
- focus : red brackets

IS: Visualization

Feigenblatt

[Die Jugendlichen in Zossen] [wollen ein Musikcafé] . Das forderten [sie] [bei der ersten Zossener Runde am Dienstagabend] . new [die Politiker der Stadt] dafür Verständnis haben , ist loblich . [Mit dem Treffen im Rathaus] [ist somit auch ein Dialog zwischen den Generationen angestoßen] . Dass [die beiden geladenen Jugendlichen] [im Laufe des Abends] [immer weniger zu Wort kamen , war sicher keine böse Absicht , ärgerlich ist es trotzdem] . [Und aberwitzig dazu] . Es entbehrt nicht der Komik , wenn sich [drei Erwachsene - Karola Andrae (Bürgerbündnis/FDP) , Susanne Michler (CDU) und Joachim Zanow (SPD) -] darüber streiten , was Jugendliche wollen und brauchen , ohne auf die Idee zu kommen , sie selbst zu fragen . Und das , obwohl [sie] ihnen gegenüber sitzen . [Die Jugendlichen] [wurden somit zum bloßen Feigenblatt degradiert] . [Nicht über sondern mit ihnen hätten] [die Politiker] [reden sollen] . Damit ist eine große Chance vertan . Vielleicht klappt es [bei der nächsten Runde Anfang 2002 .] [Dann] werden auch mehr Jugendliche eingeladen . [In der Gruppe können] [sie] [sich hoffentlich mehr Gehör verschaffen] . [Vielleicht finden dann auch] [Vertreter von PDS und Gewerbeverein] [ihren Weg ins Rathaus] . [Die] [glänzten diesmal noch mit Abwesenheit] .

- information status: **new**

IS: Visualization

Feigenblatt

[Die Jugendlichen in Zossen] [wollen ein Musikcafé] . Das forderten [sie] [bei der ersten Zossener Runde am Dienstagabend] . Dass [die Politiker der Stadt] dafür Verstärkung geben, ist loblich . [Mit dem Treffen im Rathaus] [ist somit auch ein Dialog zwischen den Generationen angestoßen] . Dass [die beiden geladenen Jugendlichen] [im Laufe des Abends] [immer weniger zu Wort kamen, war sicher keine böse Absicht, ärgerlich ist es trotzdem] . [Und aberwitzig dazu] . Es entbehrt nicht der Komik, wenn sich [drei Erwachsene - Karola Andrae (Bürgerbündnis/FDP) , Susanne Michler (CDU) und Joachim Zanow (SPD) -] darüber streiten, was Jugendliche wollen und brauchen, ohne auf die Idee zu kommen, sie selbst zu fragen . Und das, obwohl [sie] ihnen gegenüber sitzen . [Die Jugendlichen] [wurden somit zum bloßen Feigenblatt degradiert] . [Nicht über sondern mit ihnen hätten] [die Politiker] [reden sollen] . Damit ist eine große Chance vertan . Vielleicht klappt es [bei der nächsten Runde Anfang 2002 .] [Dann] werden auch mehr Jugendliche eingeladen . [In der Gruppe können] [sie] [sich hoffentlich mehr Gehör verschaffen] . [Vielleicht finden dann auch] [Vertreter von PDS und Gewerbeverein] [ihren Weg ins Rathaus] . [Die] [glänzten diesmal noch mit Abwesenheit] .

- information status: given (active, inactive)

IS: Visualization

Feigenblatt

[Die Jugendlichen in Zossen] [wollen ein Musikcafé] . Das forderten [sie] [bei der ersten Zossener Runde am Dienstagabend] . Dass [die Politiker der Stadt] dafür Verständnis haben , ist löblich . [Mit dem Treffen im Rathaus] [ist somit auch ein Dialog zwischen den Generationen angestoßen] . Dass [die beiden geladenen Jugendlichen] [im Laufe des Abends] [immer weniger zu Wort kamen , war sicher keine böse Absicht , ärgerlich ist es trotzdem] . [Und aberwitzig dazu] . giv-inactivert nicht der Komik , wenn sich [drei Erwachsene - Karola Andrae (Bürgerbündnis/FDP) , Susanne Michler (CDU) und Joachim Zanow (SPD) -] darüber streiten , was Jugendliche wollen und brauchen , ohne auf die Idee zu kommen , sie selbst zu fragen . Und das , obwohl [sie] ihnen gegenüber sitzen . [Die Jugendlichen] [wurden somit zum bloßen Feigenblatt degradiert] . [Nicht über sondern mit ihnen hätten] [die Politiker] [reden sollen] . Damit ist eine große Chance vertan . Vielleicht klappt es [bei der nächsten Runde Anfang 2002 .] [Dann] werden auch mehr Jugendliche eingeladen . [In der Gruppe können] [sie] [sich hoffentlich mehr Gehör verschaffen] . [Vielleicht finden dann auch] [Vertreter von PDS und Gewerbeverein] [ihren Weg ins Rathaus] . [Die] [glänzten diesmal noch mit Abwesenheit] .

- information status: given (active, inactive)

IS: Visualization

Feigenblatt

[Die Jugendlichen in Zossen] [wollen ein Musikcafé] . Das forderten [sie] [bei der ersten Zossener Runde am Dienstagabend] . Dass [die Politiker der Stadt] dafür Verständnis haben , ist löblich . [Mit dem Treffen im Rathaus] [ist somit auch ein Dialog zwischen den Generationen angestoßen] . Dass [die beiden geladenen Jugendlichen] [im Laufe des Abends] [immer weniger zu Wort kamen , war sicher keine böse Absicht , ärgerlich ist es trotzdem] . [Und aberwitzig dazu] . Es entbehrt nicht der Komik , wenn sich [drei Erwachsene - Karola Andrae (Bürgerbündnis/FDP) , Susanne Michler (CDU) und Joachim Zanow (SPD) -] darüber streiten , was Jugendliche wollen und brauchen , ohne auf die Idee zu kommen , sie selbst zu fragen . Und das , obwohl [sie] ihnen gegenüber sitzen . [Die Jugendlichen] [wurden somit zum bloßen Feigenblatt degradiert] . [Nicht über sondern mit ihnen hätten] [die Politiker] [reden sollen] . Damit ist eine große Chance vertan . Vielleicht klappt es [bei der nächsten Runde Anfang 2002 .] [Dann] werden auch mehr Jugendliche eingeladen . [In der Gruppe können] [sie] [sich hoffentlich mehr Gehör verschaffen] . [Vielleicht finden dann auch] [Vertreter von PDS und Gewerbeverein] [ihren Weg ins Rathaus] . [Die] [glänzten diesmal noch mit Abwesenheit] .

- information status: **accessible** (gen, inf...)

IS: Visualization

Feigenblatt

[Die Jugendlichen in Zossen] [wollen ein Musikcafé] . Das forderten [sie] [bei der ersten Zossener Runde am Dienstagabend] . Dass [die Politiker der Stadt] dafür Verständnis haben , ist löblich . [Mit dem Treffen im Rathaus] [ist somit auch ein Dialog zwischen den Generationen angestoßen] . Dass [die beiden geladenen Jugendlichen] [im Laufe des Abends] [immer wenn acc-inf Wort kamen , war sicher keine böse Absicht , ärgerlich ist es trotzdem] . [Und aberwitzig dazu] . Es entbehrt nicht der Komik , wenn sich [drei Erwachsene - Karola Andrae (Bürgerbündnis/FDP) , Susanne Michler (CDU) und Joachim Zanow (SPD) -] darüber streiten , was Jugendliche wollen und brauchen , ohne auf die Idee zu kommen , sie selbst zu fragen . Und das , obwohl [sie] ihnen gegenüber sitzen . [Die Jugendlichen] [wurden somit zum bloßen Feigenblatt degradiert] . [Nicht über sondern mit ihnen hätten] [die Politiker] [reden sollen] . Damit ist eine große Chance vertan . Vielleicht klappt es [bei der nächsten Runde Anfang 2002 .] [Dann] werden auch mehr Jugendliche eingeladen . [In der Gruppe können] [sie] [sich hoffentlich mehr Gehör verschaffen] . [Vielleicht finden dann auch] [Vertreter von PDS und Gewerbeverein] [ihren Weg ins Rathaus] . [Die] [glänzten diesmal noch mit Abwesenheit] .

- information status: **accessible** (gen, inf...)

IS: Visualization

Feigenblatt

[Die Jugendlichen in Zossen] [wollen ein Musikcafé] . Das forderten [sie] [bei der ersten Zossener Runde am Dienstagabend] . Dass [die Politiker der Stadt] dafür Verständnis haben , ist löblich . [Mit dem Treffen im Rathaus] [ist somit auch ein Dialog zwischen den Generationen angestoßen] . Dass [die beiden geladenen Jugendlichen] [im Laufe des Abends] [immer weniger zu Wort kamen , war sicher keine böse Absicht , ärgerlich ist es trotzdem] . [Und aberwitzig dazu] . Es entbehrt nicht der Komik , wenn sich [drei Erwachsene - Karo ^{idiom} Trae (Bürgerbündnis/FDP) , Susanne Michler (CDU) und Joachim Zanow (SPD) -] darüber streiten , was Jugendliche wollen und brauchen , ohne auf die Idee zu kommen , sie selbst zu fragen . Und das , obwohl [sie] ihnen gegenüber sitzen . [Die Jugendlichen] [wurden somit zum bloßen Feigenblatt degradiert] . [Nicht über sondern mit ihnen hätten] [die Politiker] [reden sollen] . Damit ist eine große Chance vertan . Vielleicht klappt es [bei der nächsten Runde Anfang 2002 .] [Dann] werden auch mehr Jugendliche eingeladen . [In der Gruppe können] [sie] [sich hoffentlich mehr Gehör verschaffen] . [Vielleicht finden dann auch] [Vertreter von PDS und Gewerbeverein] [ihren Weg ins Rathaus] . [Die] [glänzten diesmal noch mit Abwesenheit] .

- information status: idiom

IS: Visualization

Feigenblatt

[Die Jugendlichen in Zossen] [wollen ein Musikcafé] . Das forderten [sie] [bei der ersten Zossener Runde am Dienstagabend] . Dass [die Politiker der Stadt] dafür Verständnis haben , ist löblich . [Mit dem Treffen im Rathaus] [ist somit auch ein Dialog zwischen den Generationen angestoßen] . Dass [die beiden geladenen Jugendlichen] [im Laufe des Abends] [immer weniger zu Wort kamen , war sicher keine böse Absicht , ärgerlich ist es trotzdem] . [Und aberwitzig dazu] . Es entbehrt nicht der Komik , wenn sich [drei Erwachsene - Karo ^{idiom} Trae (Bürgerbündnis/FDP) , Susanne Michler (CDU) und Joachim Zanow (SPD) -] darüber streiten , was Jugendliche wollen und brauchen , ohne auf die Idee zu kommen , sie selbst zu fragen . Und das , obwohl [sie] ihnen gegenüber sitzen . [Die Jugendlichen] [wurden somit zum bloßen Feigenblatt degradiert] . [Nicht über sondern mit ihnen hätten] [die Politiker] [reden sollen] . Damit ist eine große Chance vertan . Vielleicht klappt es [bei der nächsten Runde Anfang 2002 .] [Dann] werden auch mehr Jugendliche eingeladen . [In der Gruppe können] [sie] [sich hoffentlich mehr Gehör verschaffen] . [Vielleicht finden dann auch] [Vertreter von PDS und Gewerbeverein] [ihren Weg ins Rathaus] . [Die] [glänzten diesmal noch mit Abwesenheit] .

- topic – blue brackets: [frame-setter], [aboutness]
- [focus]

IS: Query Examples

ab-topic with given and new information

```
Topic="ab" & Inf-Stat="new" & Inf-Stat=/giv.*/ &  
#1 _i_ #2 & #1 _i_ #3
```

Dass [die Politiker der Stadt] dafür Verständnis haben,
ist loblich.

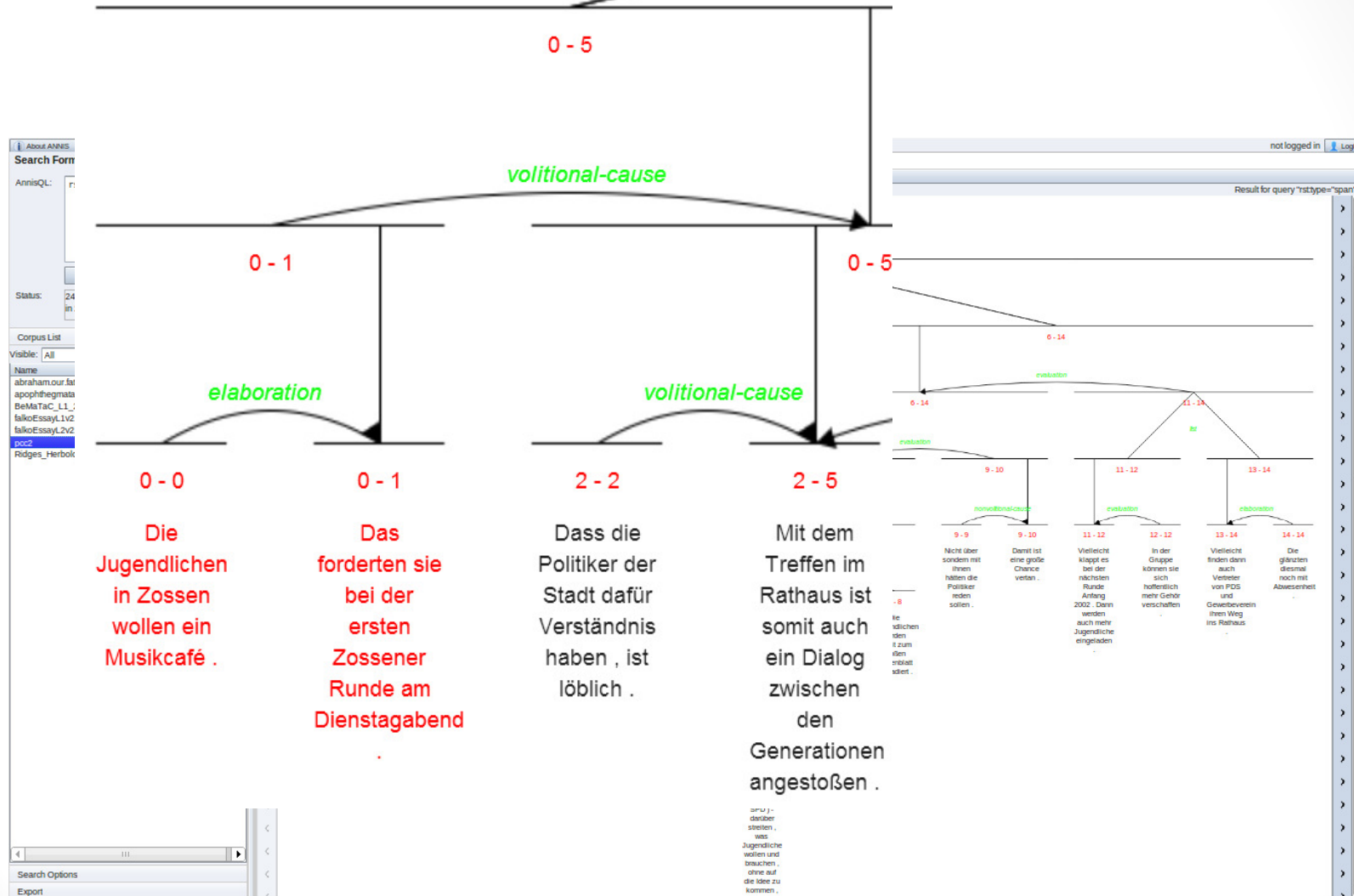
IS: Query Examples

ab- and fs-topic in the same sentence, ab precedes fs

```
cat="S" & Topic="ab" & Topic="fs" &  
#1 _i_ #2 & #1 _i_ #3 & #2 .* #3
```

Dass [die beiden geladenen Jugendlichen]
[im Laufe des Abends] immer weniger zu Wort kamen,
war sicher keine böse Absicht.[...]

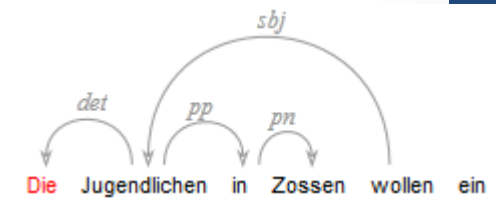
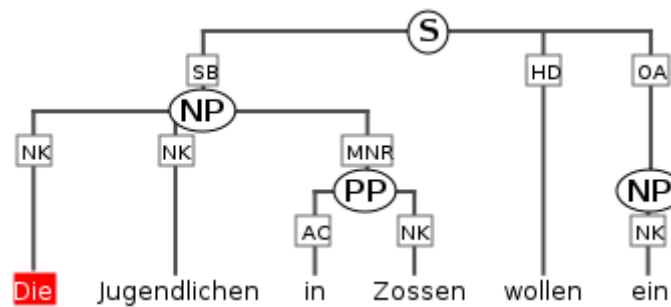
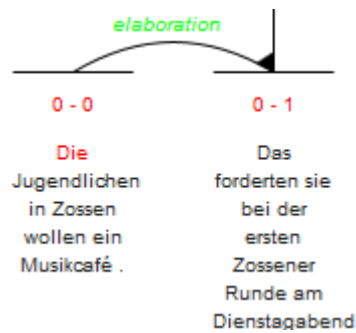
RST: Visualization



Data integration in ANNIS

- Complex corpora require complex workflows
- Need to deal with different types of information
- A data "pluriverse"

Data pluriverse



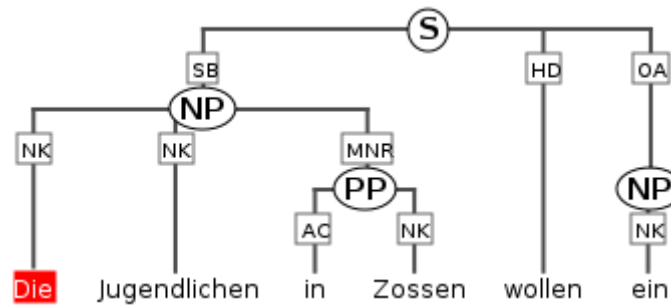
Focus_newInf						nf-unsol		
Inf-Stat		new			new		new	
NP		NP			NP		NP	
PP				PP				
Sent		s						
Topic		ab						
heading	heading							
tok	Feigenblatt	Die	Jugendlichen	in	Zossen	wollen	ein	

Data pluriverse

elaboration

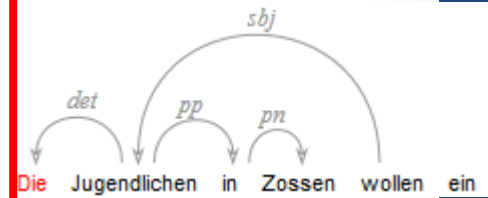
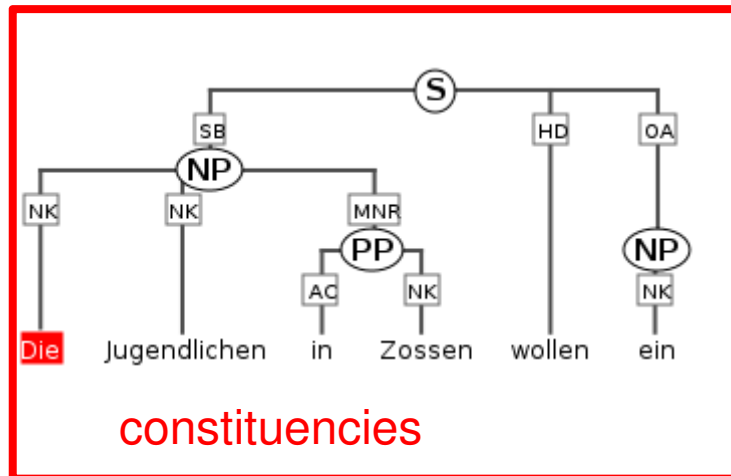
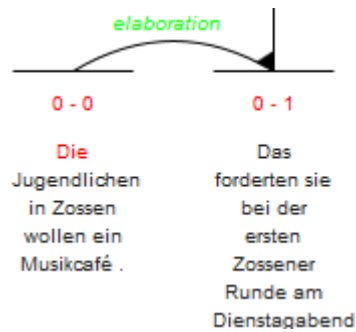
0 - 0	0 - 1
<p style="color: red;">Die</p> <p>Jugendlichen in Zossen wollen ein Musikcafé .</p>	<p>Das</p> <p>forderten sie bei der ersten Zossen Runde am Dienstagabend</p>

rhetorical structure



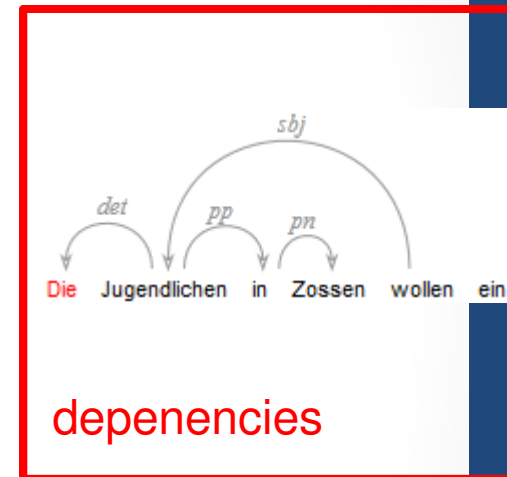
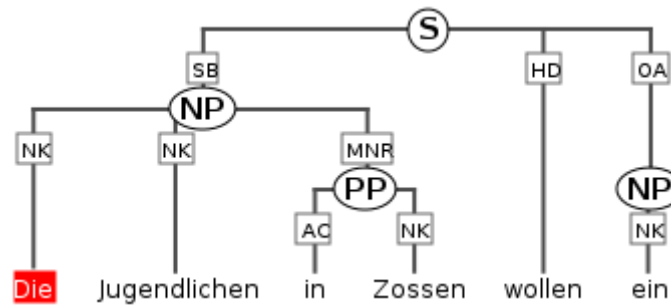
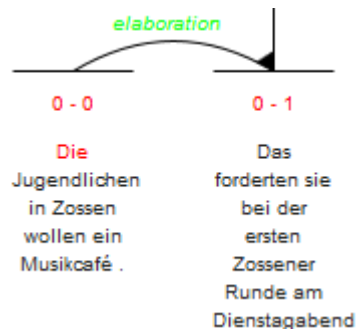
Focus_newInf						nf-unsol		
Inf-Stat		new			new		new	
NP		NP			NP		NP	
PP				PP				
Sent		s						
Topic		ab						
heading	heading							
tok	Feigenblatt	Die	Jugendlichen	in	Zossen	wollen	ein	

Data pluriverse



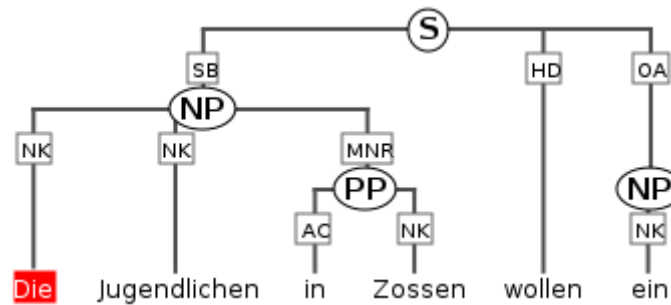
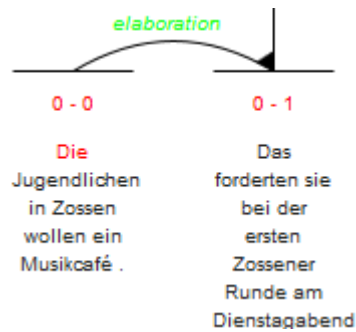
Focus_newInf						nf-unsol	
Inf-Stat		new			new		new
NP		NP			NP		NP
PP				PP			
Sent		s					
Topic		ab					
heading	heading						
tok	Feigenblatt	Die	Jugendlichen	in	Zossen	wollen	ein

Data pluriverse



Focus_newInf						nf-unsol		
Inf-Stat		new			new		new	
NP		NP			NP		NP	
PP				PP				
Sent		s						
Topic		ab						
heading	heading							
tok	Feigenblatt	Die	Jugendlichen	in	Zossen	wollen	ein	

Data pluriverse



Focus_newInf						nf-unsol		
Inf-Stat		new			new		new	
NP		NP			NP		NP	
PP				PP				
Sent		s						
Topic		ab						
heading	heading							
tok	Feigenblatt	Die	Jugendlichen	in	Zossen	wollen	ein	

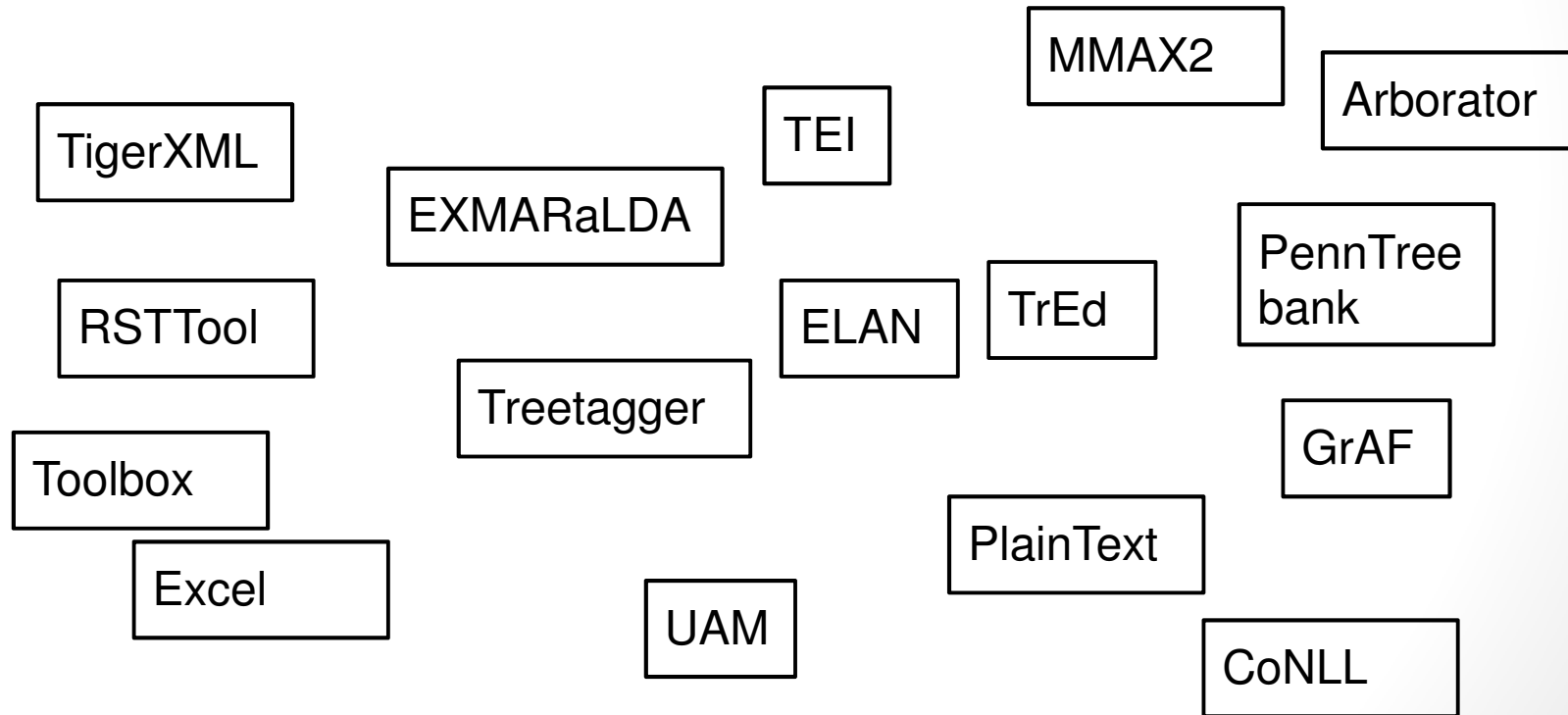
information structure

But how to bring them into ANNIS?

- Corpora aren't born in ANNIS
- They come from:

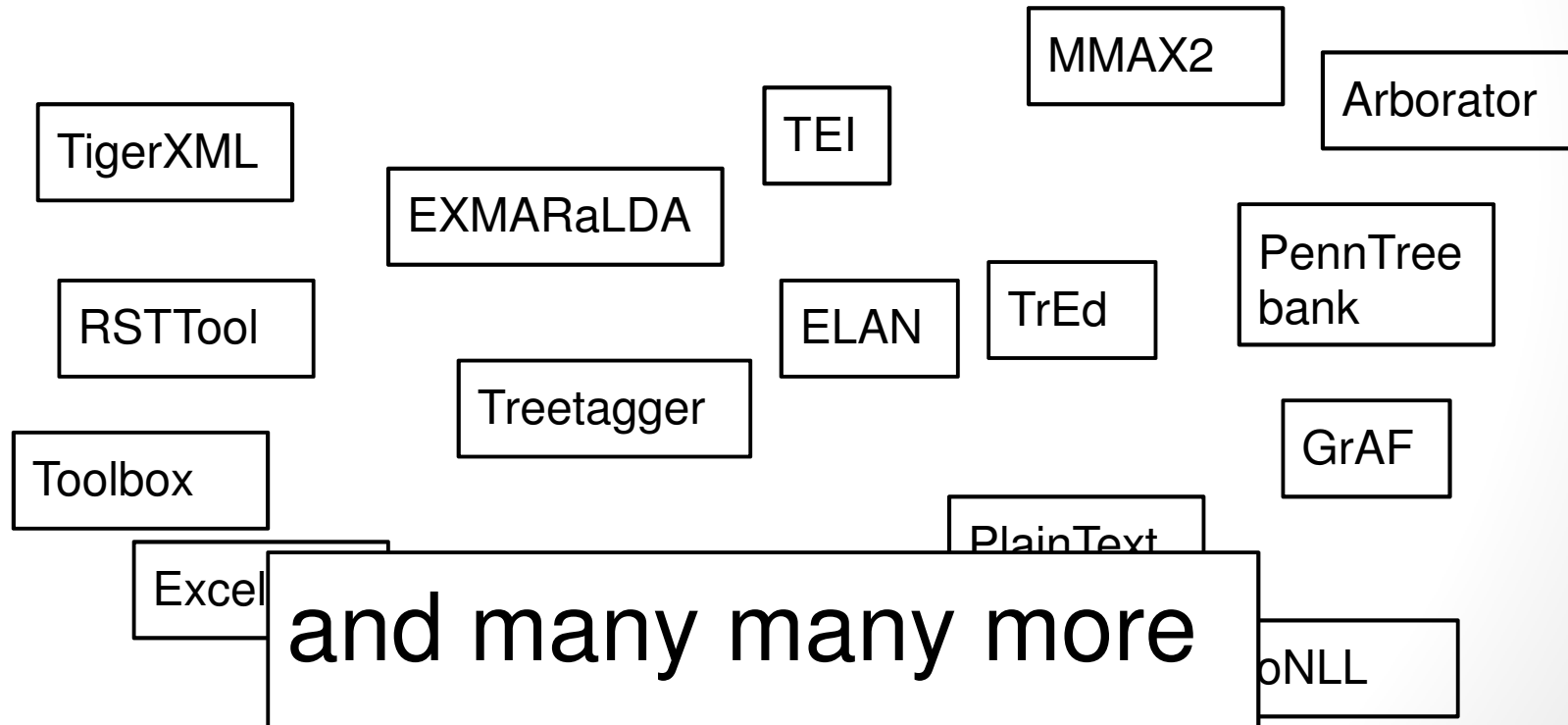
But how to bring them into ANNIS?

- Corpora aren't born in ANNIS
- They come from:



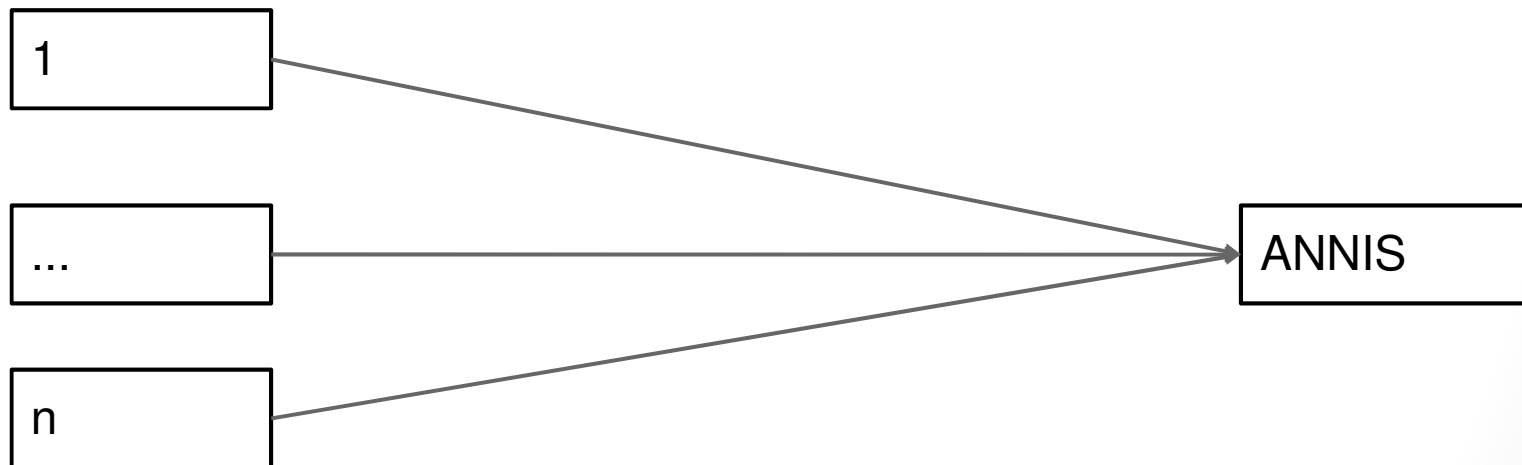
But how to bring them into ANNIS?

- Corpora aren't born in ANNIS
- They come from:



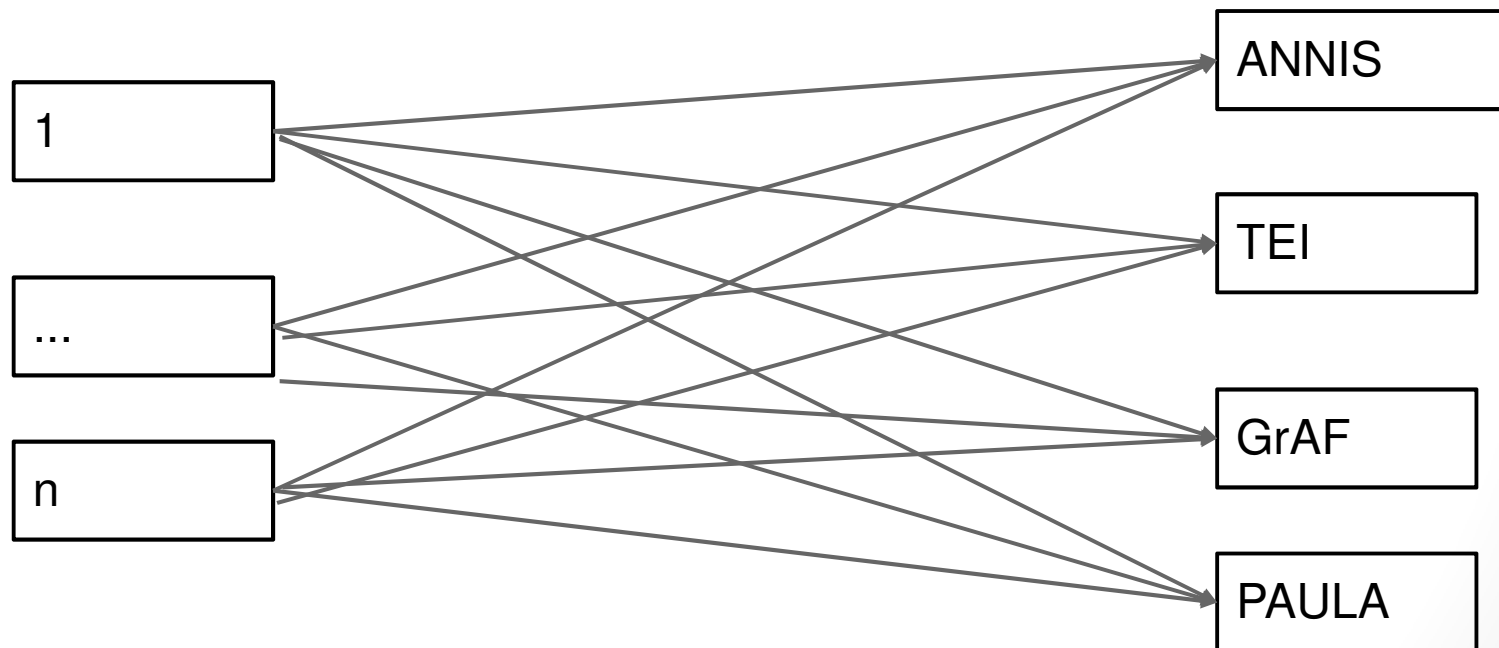
Bringing them into ANNIS

- Needs n mappings for n formats



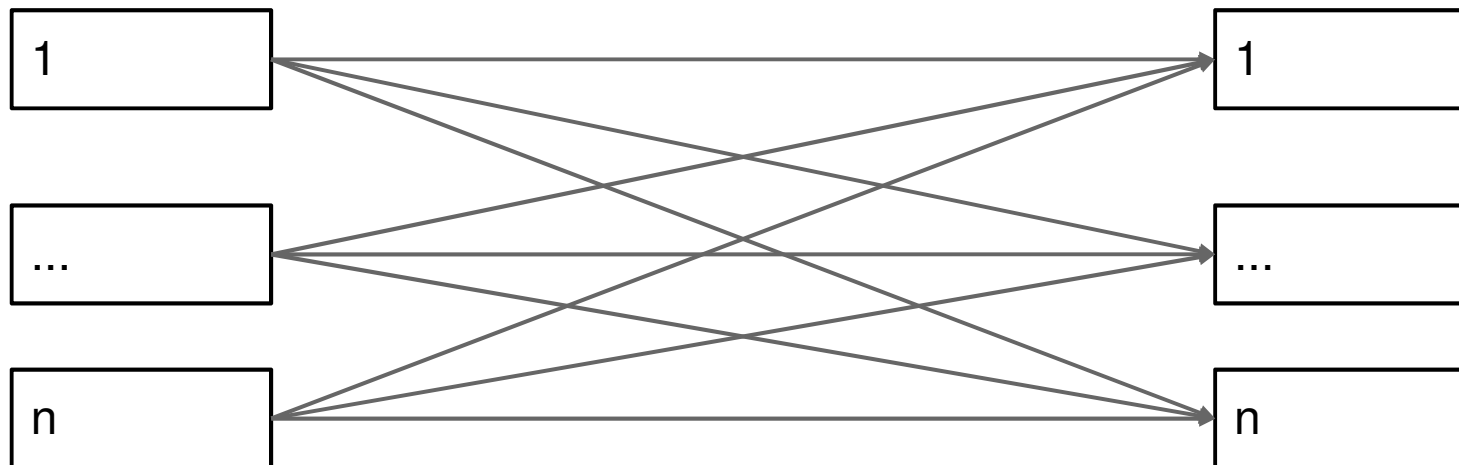
But not only into ANNIS

- corpora are expensive, we want to produce sustainable data



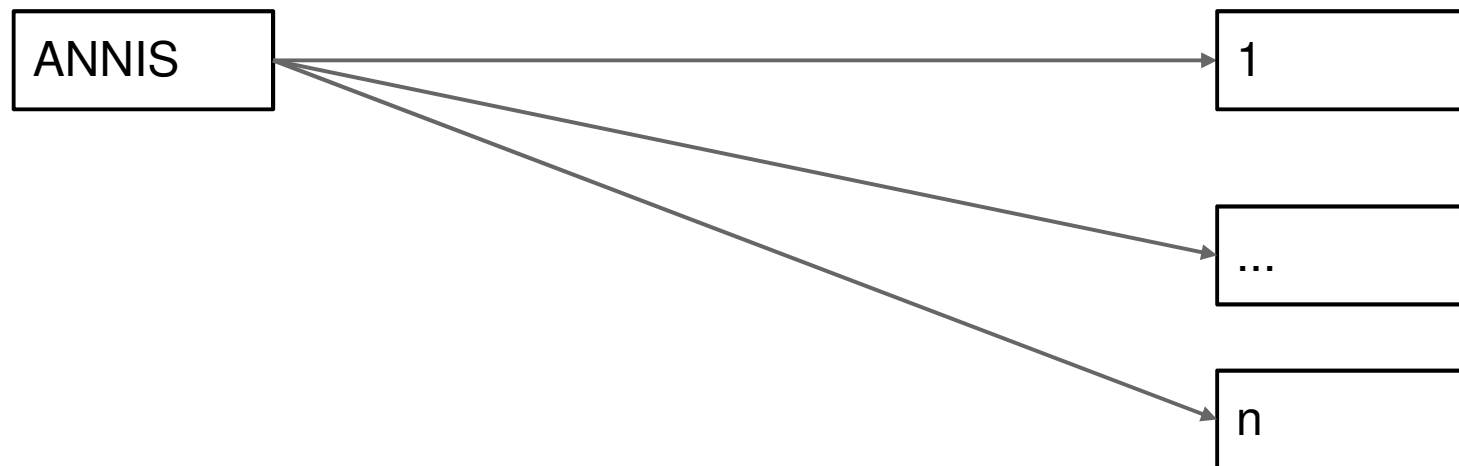
But not only into ANNIS

- We want to enhance data for further phenomena



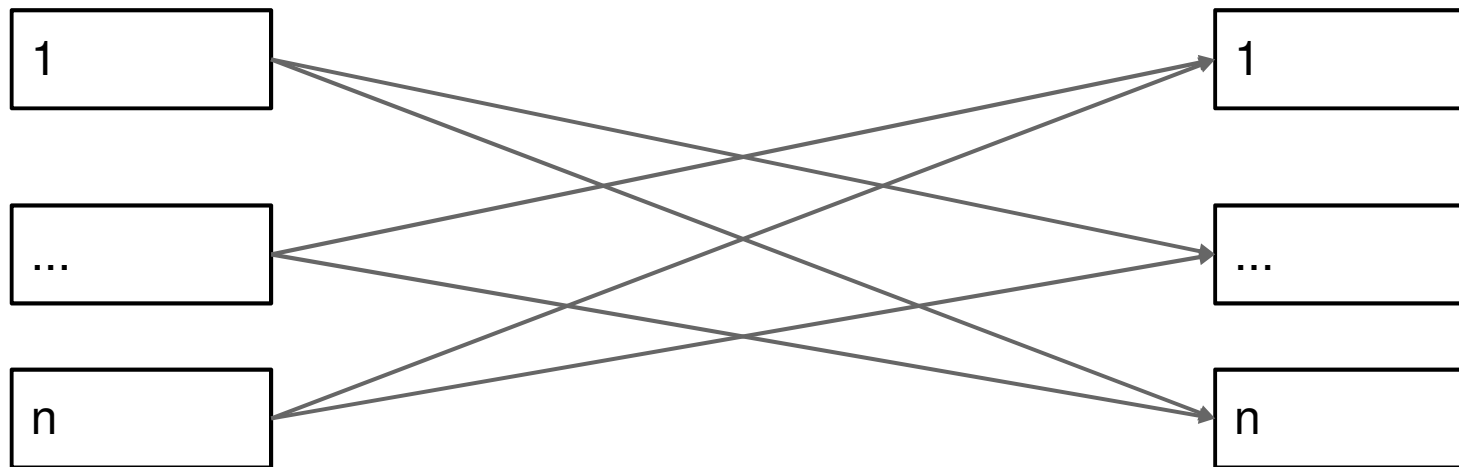
How to get them out of ANNIS

- we want to export and extend data from ANNIS



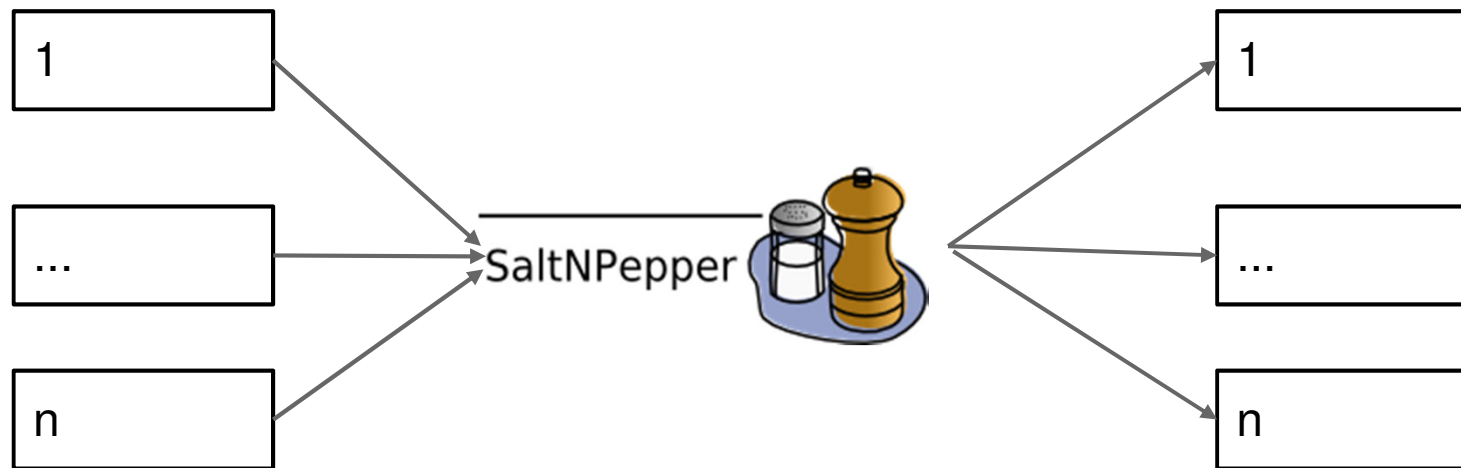
Conversion task

- for n formats: n^2 - n mappings :-)

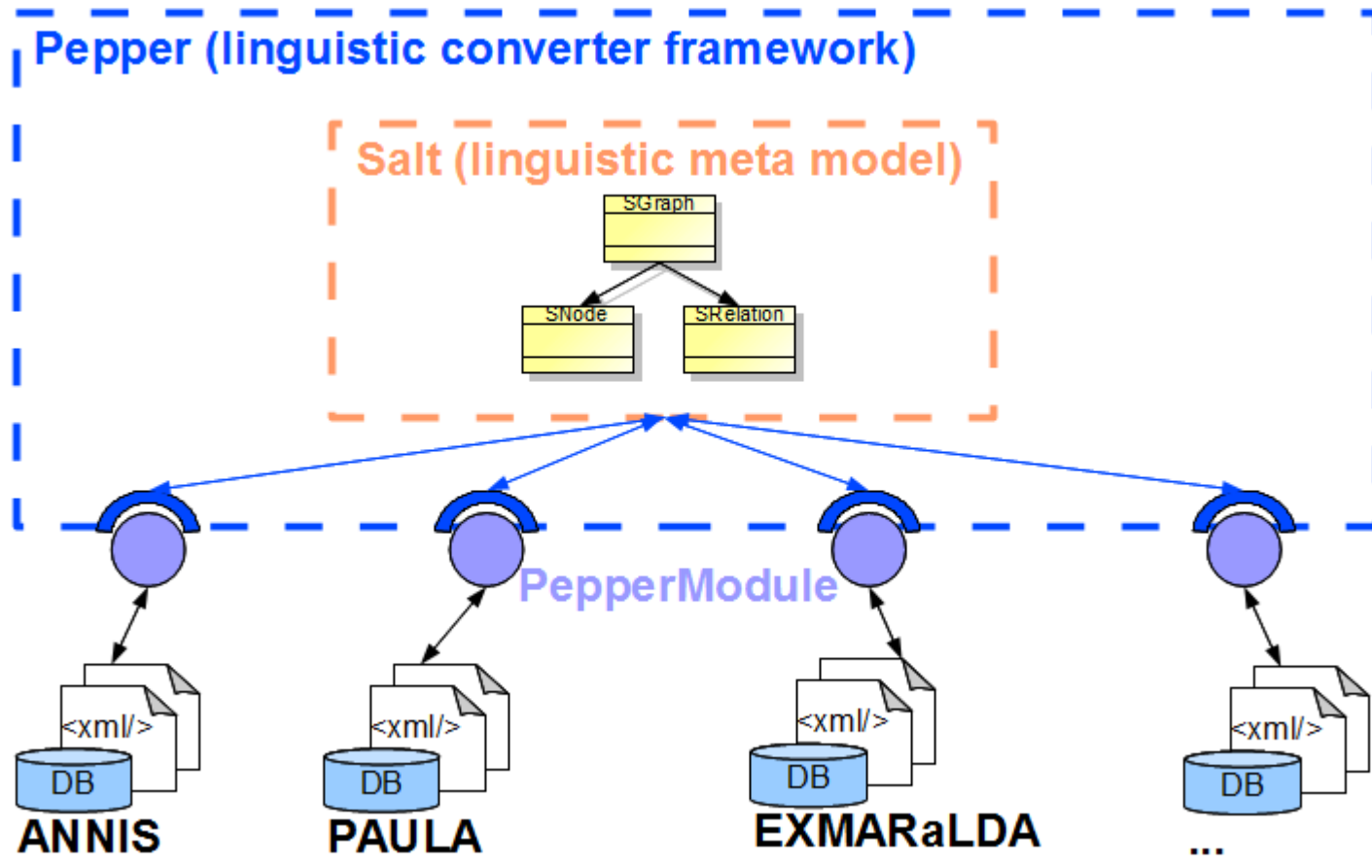


Conversion task

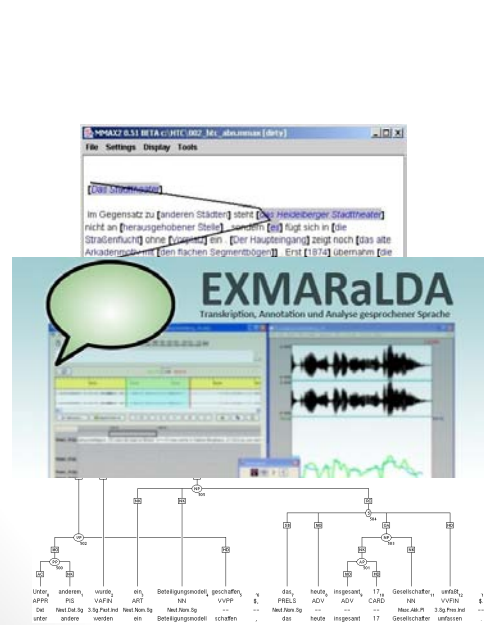
- reduces mappings to $2n$



SaltNPepper



infrastructure

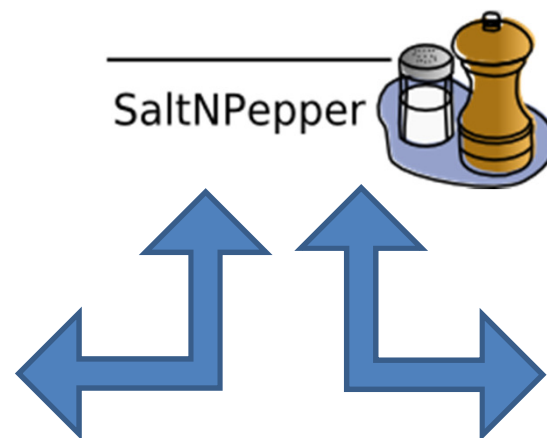


The screenshot shows the EXMARaLDA interface with a transcription window displaying German text: "Im Gegensatz zu [anderen Städten] steht [die] Heidelberger [Stadthälfte] nicht an [herausgehobener Stelle], sondern [es] fügt sich in [die Straßenflucht] ohne [Fußweg] ein. [Der Hauptgang] zeigt noch [das alte Ankerdenkmal] mit [den Büchen Segenerbögen]. Erst [1974] übernahm Ebe". Below the text, there are audio waveforms and a tree diagram representing a phrase structure grammar.

EXMARaLDA
Transkription, Assoziation und Analyse gesprochener Sprache

Unter anderem wurden ein Belegprotokoll geschaffen
appet mit vierzig
Die Neu-Geb. 3.0y.Fak.Ist.wei.ken.ig
Unter andere werden ein Belegprotokoll schaffen

das heute insgesamt 17 Gedächtnis umfassen



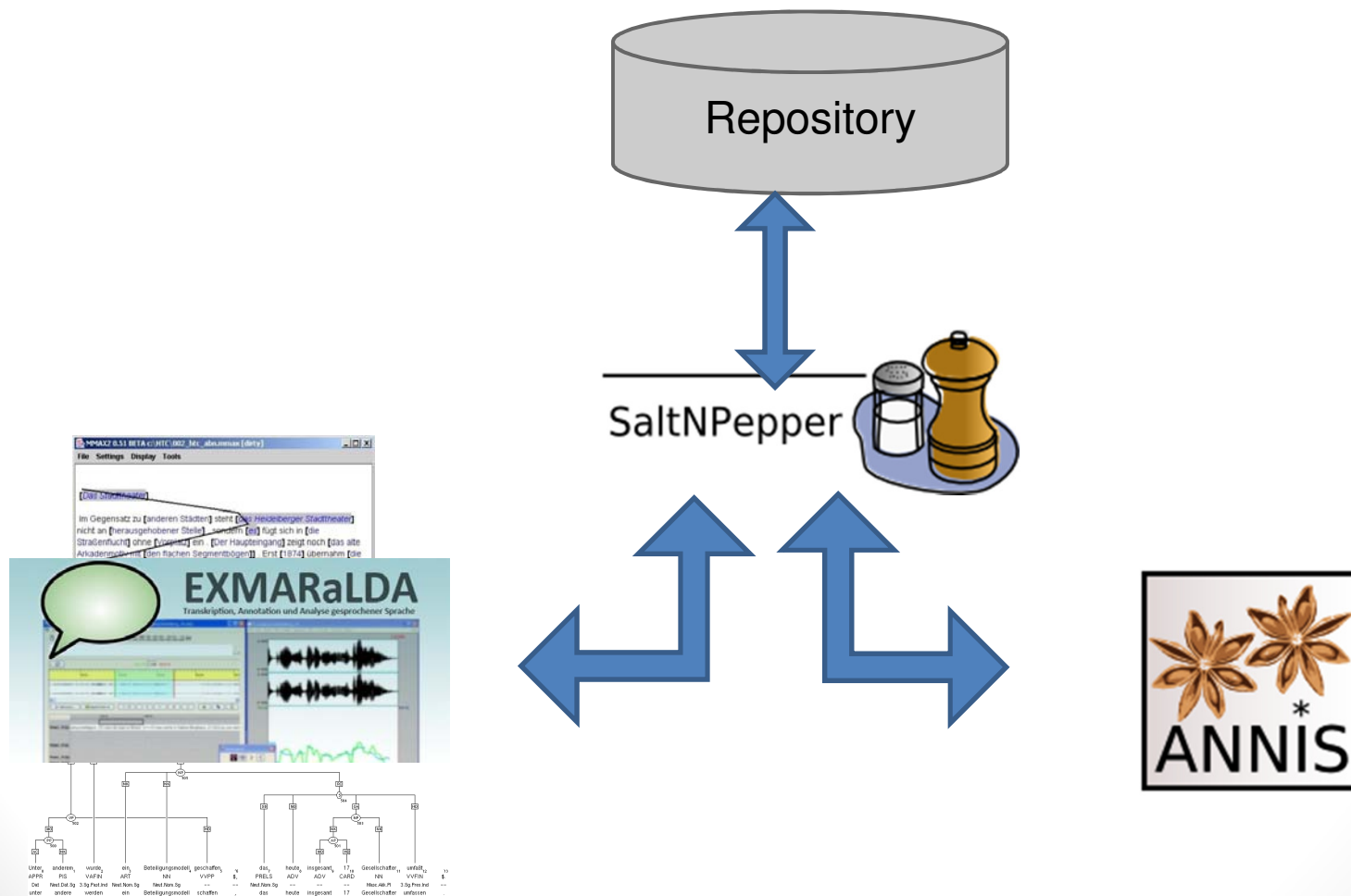
SFB infrastructure meeting

- in Goettingen april 2013
 - all sfb projects (even physics, economics ...)
- DFG:
 - each SFB should have an infrastructure project
 - sustainability of data is prerequisite for funding
 - repository center is task of discipline not of DFG

what else do we need?

- data storage
- long-term archive
- collaborative work / sharing
- enhancing data

big picture



EXMARaLDA
Transkription, Assoziation und Analyse gesprochenen Sprache

Im Gegensatz zu [anderen Städten] steht [die Heideberger Stadtheute] nicht an [herausgehobener Stelle], sondern [es] fügt sich in [die Straßenhüch] ohne [Fußweg] ein. [Der Hauptgang] zeigt noch [das alte Ankleiden] mit [dem flachen Segenröcke]. Erst [1874] übernahm [die]

Unternehmen, werden, ein, Betriebsrat, schaffen, ...
appt, die, nicht, 3, by, hat, bei, weil, kein, by, ...
unter, andere, werden, ein, Betriebsrat, schaffen, ...

Thanks for your attention!

...and many thanks to the ANNIS, SaltNPepper and LAUDATIO teams!

More about ANNIS at:

<http://www.sfb632.uni-potsdam.de/annis/>