Florian Zipser
Humboldt-Universität zu Berlin

SaltNPepper und das Formatpluriversum

LAUDATIO Workshop

2014-10-07

Heterogene Domäne

- Linguistische Daten und Phänomene erfordern viele Annotationsarten

Die	Jugendlichen	in	Zossen	wollen	ein	Musikcafé	.
der	jugendliche	in	Zossen	wollen	ein	Musikcafé	.
Nom.Pl.*	Nom.Pl.*	--	Dat.Sg.Neut	3.Pl.Pres.Ind	Acc.Sg.Neut	Acc.Sg.Neut	--
ART	NN	APPR	NE	VMFIN	ART	NN	$.

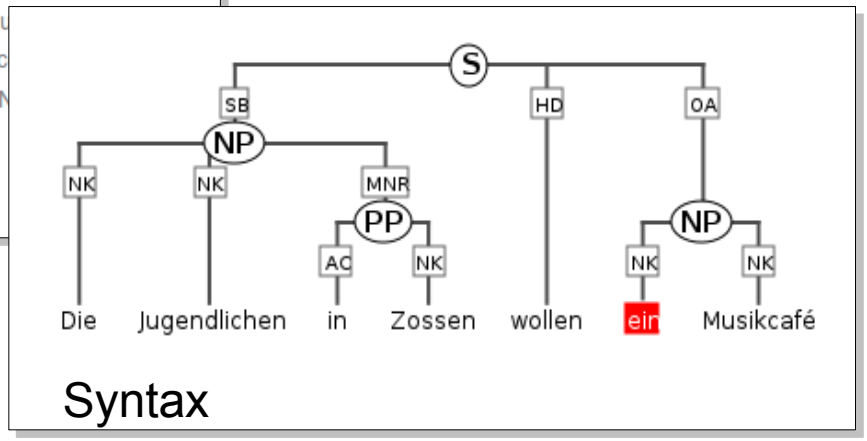
Morphologie

Heterogene Domäne

- Linguistische Daten und Phänomene erfordern viele Annotationsarten

Die	Jugendlichen	in	Zossen	wollen	ein	Musikcafé	.
der	jugendliche	in	Zossen	wollen	ein	Mu	
Nom.Pl.*	Nom.Pl.*	--	Dat.Sg.Neut	3.Pl.Pres.Ind	Acc.Sg.Neut	Ac	
ART	NN	APPR	NE	VMFIN	ART	NN	

Morphologie



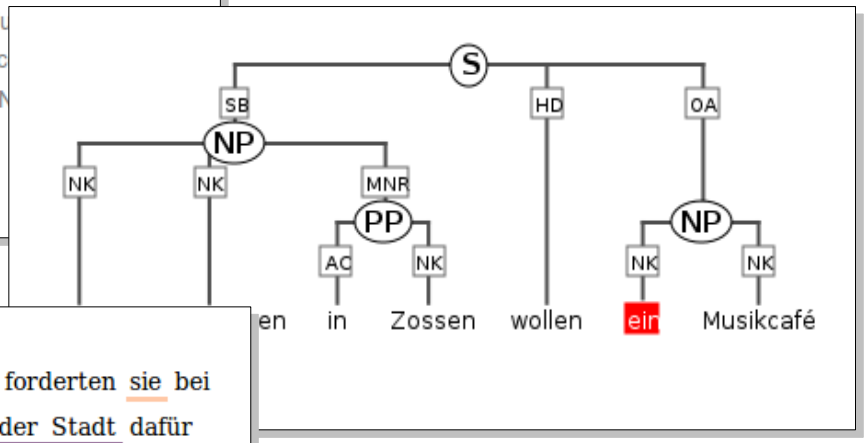
Syntax

Heterogene Domäne

- Linguistische Daten und Phänomene erfordern viele Annotationsarten

Die	Jugendlichen	in	Zossen	wollen	ein	Musikcafé	.
der	jugendliche	in	Zossen	wollen	ein	Mu	
Nom.Pl.*	Nom.Pl.*	--	Dat.Sg.Neut	3.Pl.Pres.Ind	Acc.Sg.Neut	Ac	
ART	NN	APPR	NE	VMFIN	ART	NN	

Morphologie



Feigenblatt Die Jugendlichen in Zossen wollen ein Musikcafé . Das forderten sie bei der ersten Zossener Runde am Dienstagabend . Dass die Politiker der Stadt dafür

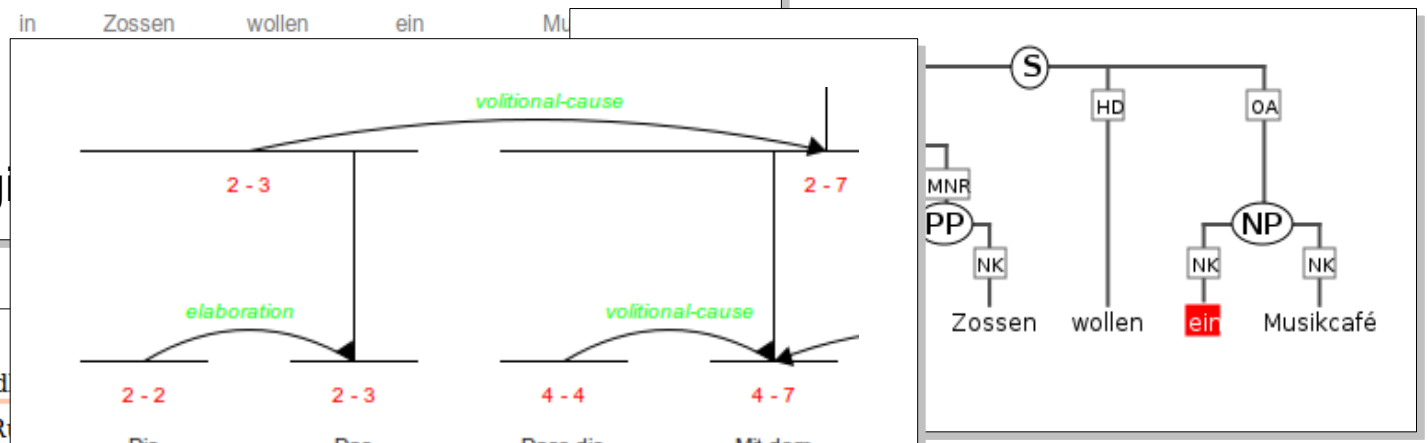
Koreferenz

Heterogene Domäne

- Linguistische Daten und Phänomene erfordern viele Annotationsarten

Die	Jugendlichen	in	Zossen	wollen	ein	Musikcafé	.
der	jugendliche	in	Zossen	wollen	ein	Mu	
Nom.Pl.*	Nom.Pl.*						
ART	NN						

Morphologi



Feigenblatt Die Jugend
der ersten Zossener R

Koreferenz

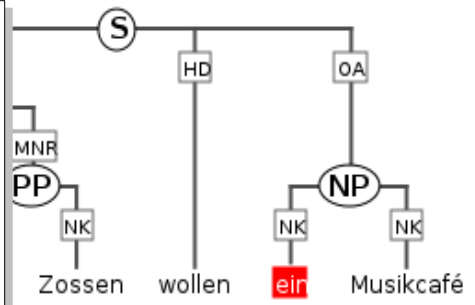
Rhetorische Strukturen

Heterogene Domäne

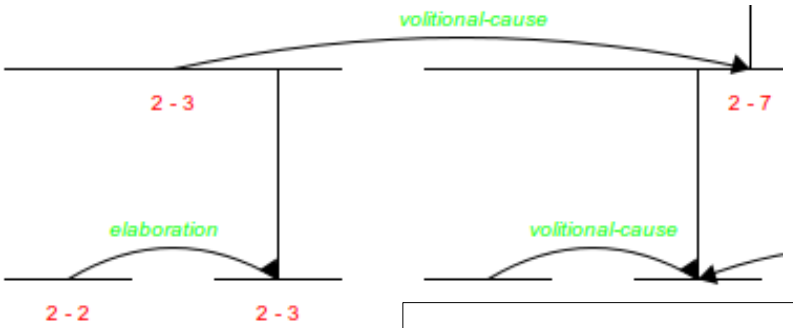
- Linguistische Daten und Phänomene erfordern viele Annotationsarten

Die	Jugendlichen	in	Zossen	wollen	ein	Musikcafé	.
der	jugendliche	in	Zossen	wollen	ein	Mu	
Nom.Pl.*	Nom.Pl.*						
ART	NN						

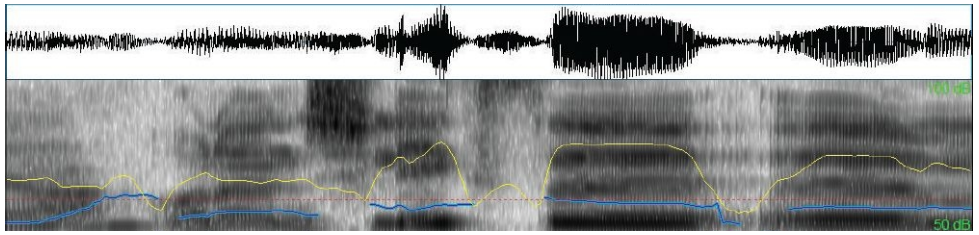
Morphologi



Feigenblatt Die Jugend
der ersten Zossener R
Koreferenz



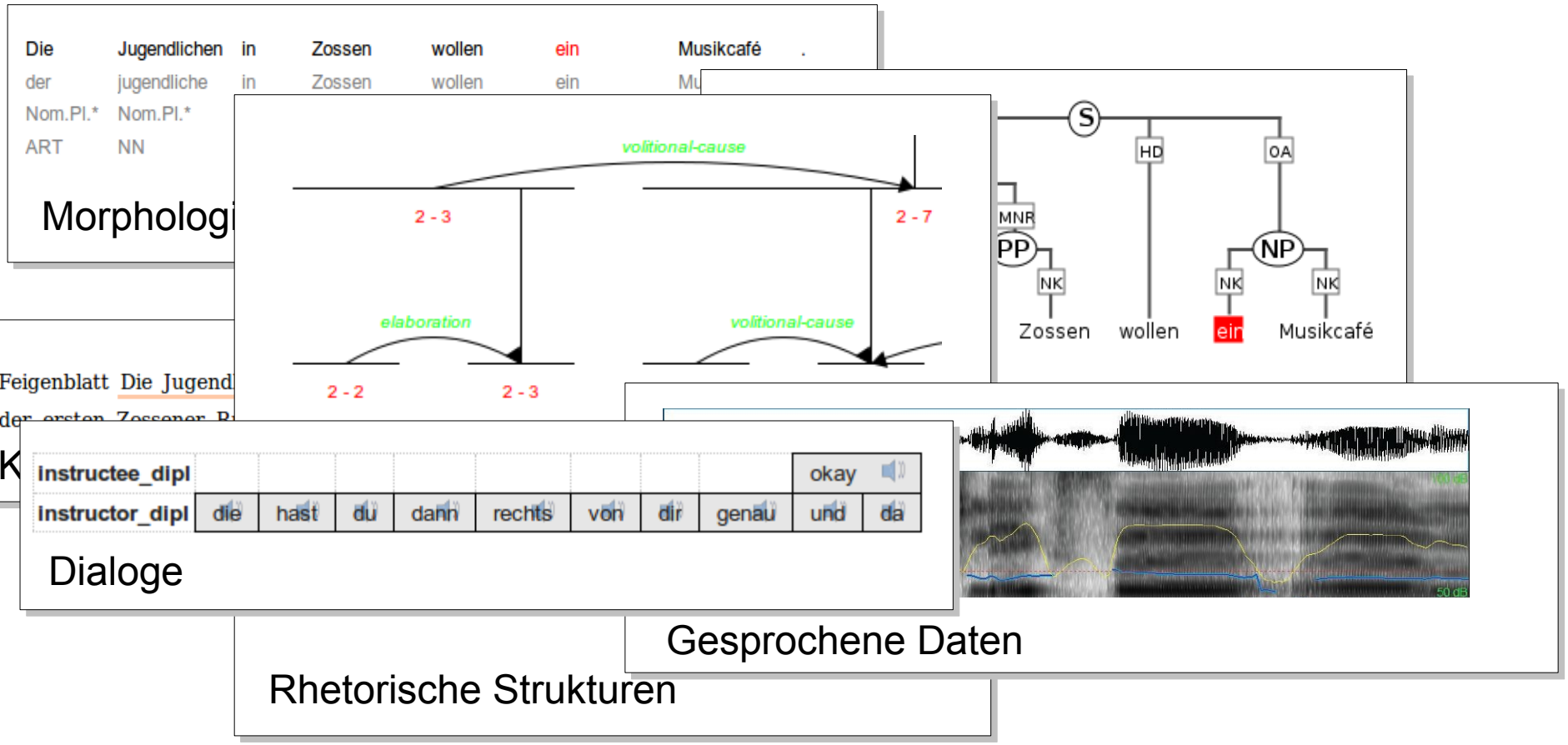
Rhetorische Strukturen



Gesprochene Daten

Heterogene Domäne

- Linguistische Daten und Phänomene erfordern viele Annotationsarten

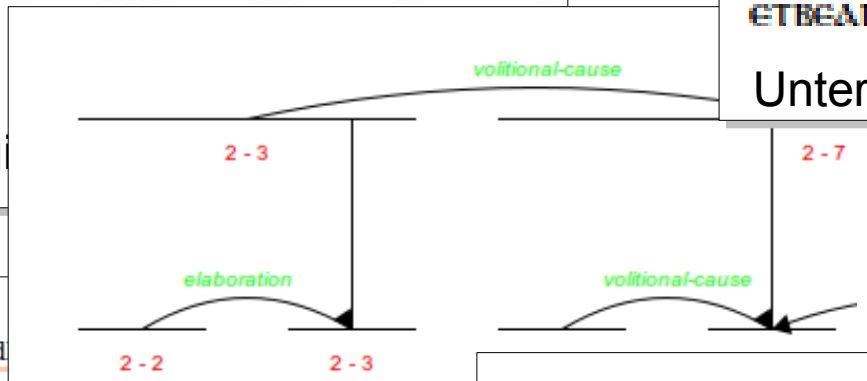


Heterogene Domäne

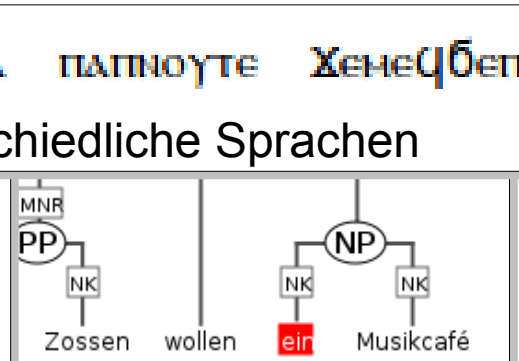
- Linguistische Daten und Phänomene erfordern viele Annotationsarten

Die	Jugendlichen	in	Zossen	wollen	ein	Musikcafé	.
der	jugendliche	in	Zossen	wollen	ein	Mu	
Nom.Pl.*	Nom.Pl.*						
ART	NN						

Morphologi



Rhetorische Strukturen



Unterschiedliche Sprachen

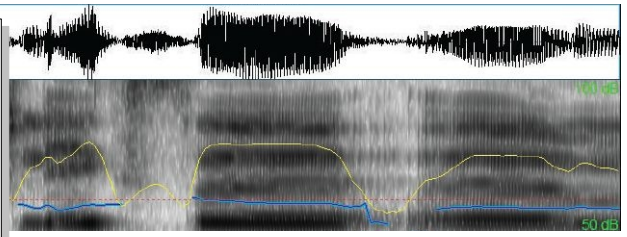
Feigenblatt Die Jugend

der ersten Zossener

K

Instructee_dipl									okay	
Instructor_dipl	die	hast	du	dann	rechts	von	dir	genau	und	da

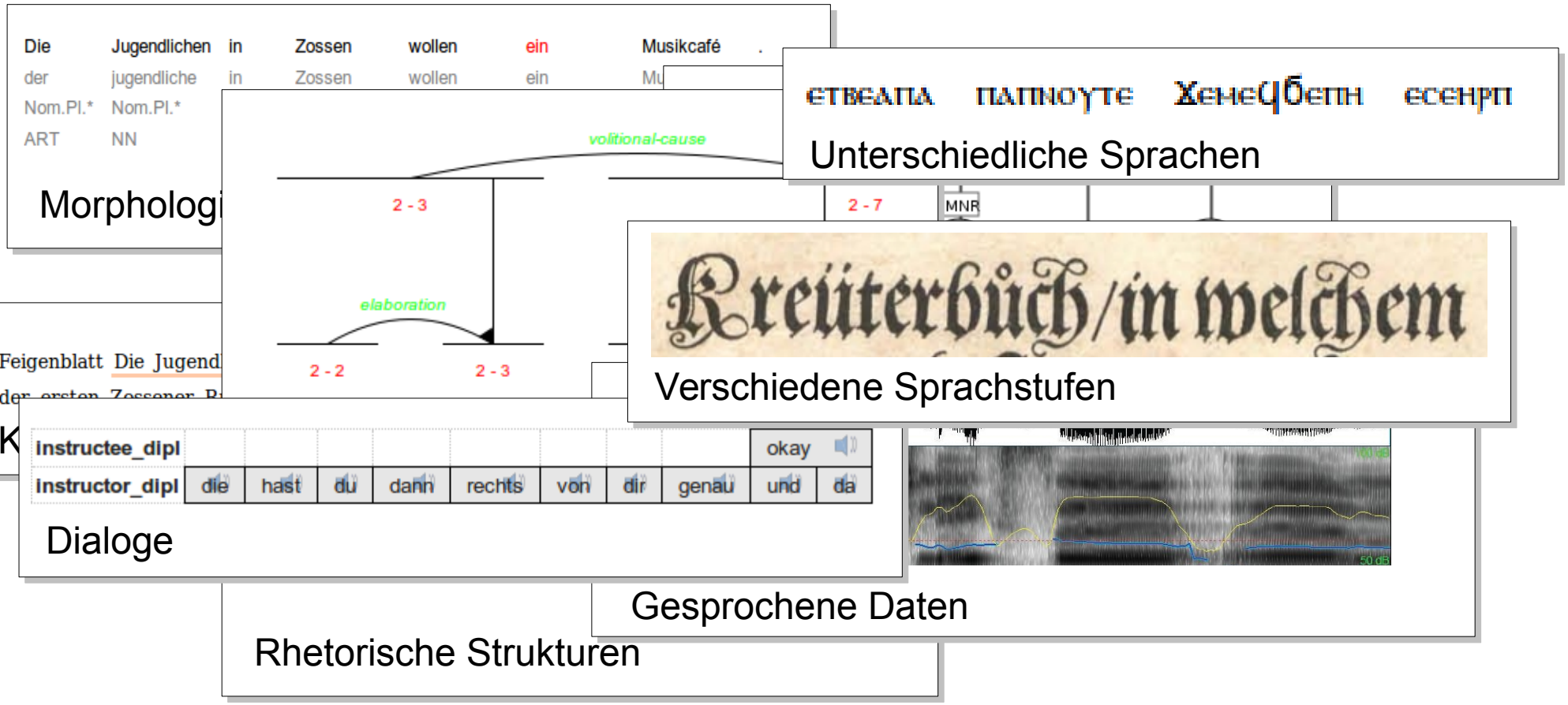
Dialoge



Gesprochene Daten

Heterogene Domäne

- Linguistische Daten und Phänomene erfordern viele Annotationsarten



Heterogene Domäne

- Viele Tools, um Daten zu bearbeiten:
  - Manuelle Annotationstools
  - semi-automatische Annotationstools
  - Automatische Annotationstools
  - Suchtools
  - Visualisierungstools

Heterogene Domäne

- Viele Tools, um Daten zu bearbeiten:

- EXMARaLDA
- Praat
- ELAN
- Tiger search
- ANNIS
- Gate
- @nnotate
- TrED
- Parser (Berkley, MALT, …)
- Arborator
- Toolbox
- Synpathie
- TreeTagger
- Weblicht
- MMAX2
- RST
- UIMA
- WebANNO
- ATOMIC
- UAM
- UIMA (dkpro, ...)
- ...

Heterogene Domäne

- Viele verschiedene Formate

GATE XML
ELAN format
MMAX2 format
TreeTagger format
TEI
Generic XML
TextGrid
GrAF
ANNIS format
TCF
Toolbox format
TigerXML
UAM format
EXMARaLDA format
PAULA
CoNLL
PTB format
rs3
UIMA CAS
Negra
MAF
PML
XLSX

Interoperabilität

- Problem 1: Interoperabilität
  - Viele Tools → gut, Nutzer können wählen
  - Aber
    - Tools können nur selten interagieren
    - Primärdaten müssen mehrmals aufbereitet werden (Tokenisierung)

Mehrebenenkorpora

- Problem 2: Mehrebenenkorpora
  - Annotation unterschiedlicher Annotationsarten (Morphologie, Syntax, Koreferenzen) erfordert defacto unterschiedliche Korpora
  - Aber: wir brauchen **ein** Korpus, das alles enthält

Nachhaltigkeit

- Problem 3: Nachhaltigkeit
  - Einige Tools werden nicht mehr weiterentwickelt
    - Formate werden nicht weiter unterstützt
    - Was ist mit den Daten???

Nachhaltigkeit

- Nachhaltigkeit der Daten erfordert Nachhaltigkeit der Speicherung
  - Im Web: HTML (W3C)
  - Allgemeine Datenbeschreibung: XML (W3C), JSON
  - Modellierung: UML/ XMI (OASIS)
  - Freitext: PDF bzw. pdf-a

Nachhaltigkeit

- Es gibt ideen zur Standardisierung:
  - TEI (TEI consortium)
  - GrAF (ISO)
  - MAF (ISO)
  - SynAF/isoTiger (ISO)

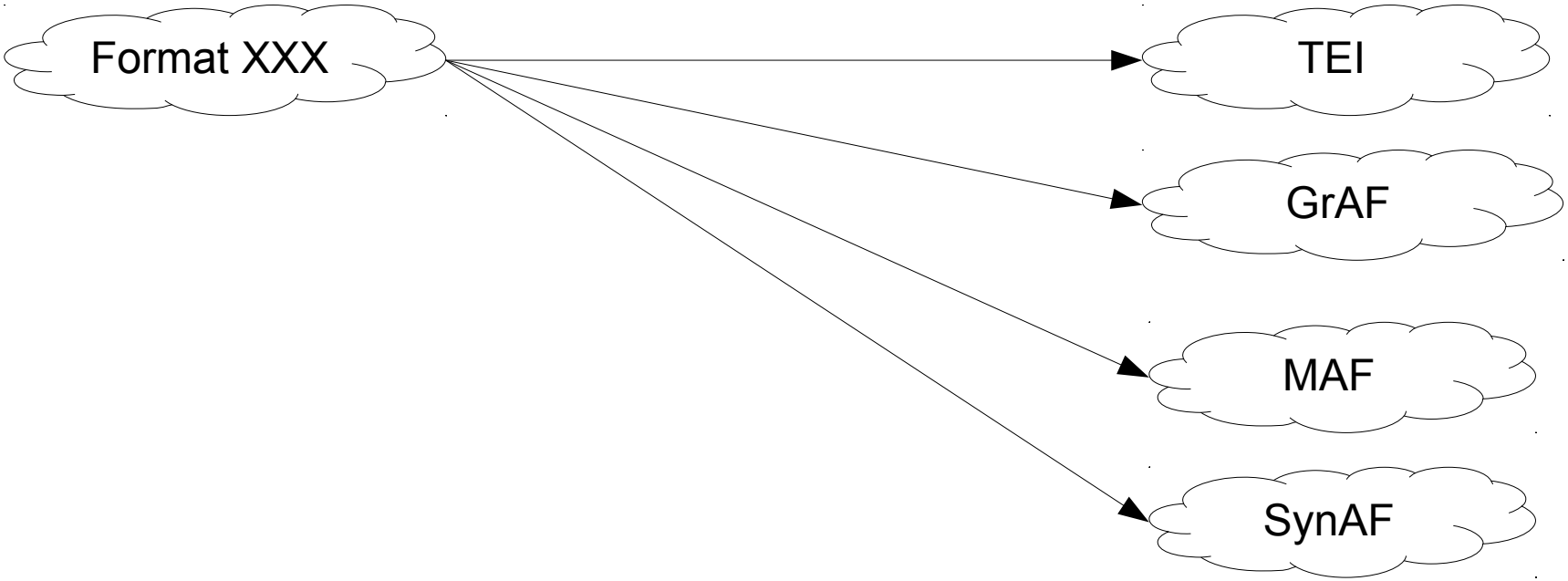
  Aber nur wenige Tools arbeiten damit, z.T.
  - Sehr komplex
  - Unausgereift
  - Standards oft jünger als Tool

Anforderungen

- Was wir brauchen:
  - Übertragung alter Daten in neue Formate/ Standards (Nachhaltigkeit)
  - Austausch der Daten zwischen unterschiedlichen Tools (Interoperabilität)
  - Verschmelzen verschiedener Annotationsarten und -ebenen (Mehrebenenkorpora)

SaltNPepper

- Nachhaltigkeit:

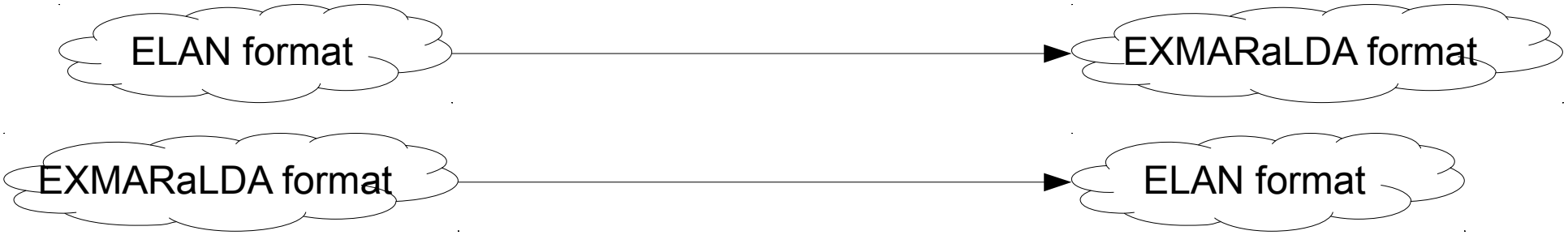


SaltNPepper

- Austausch:

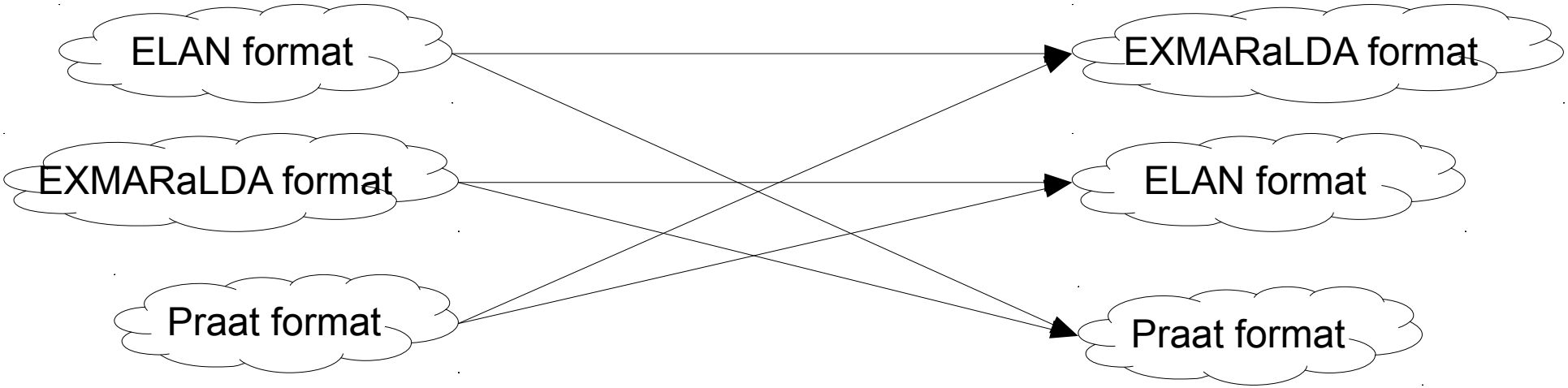
SaltNPepper

- Austausch:



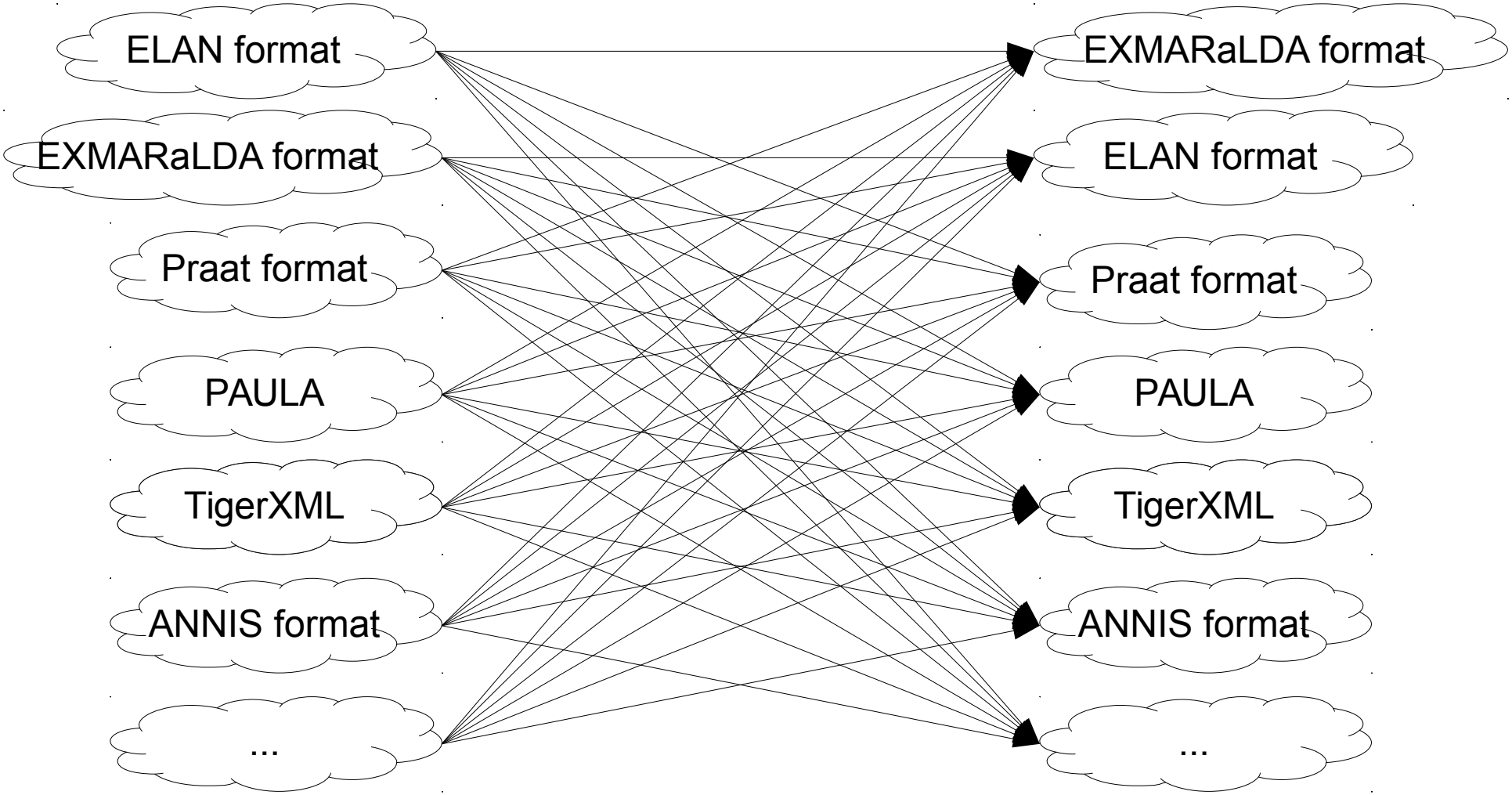
SaltNPepper

- Austausch:



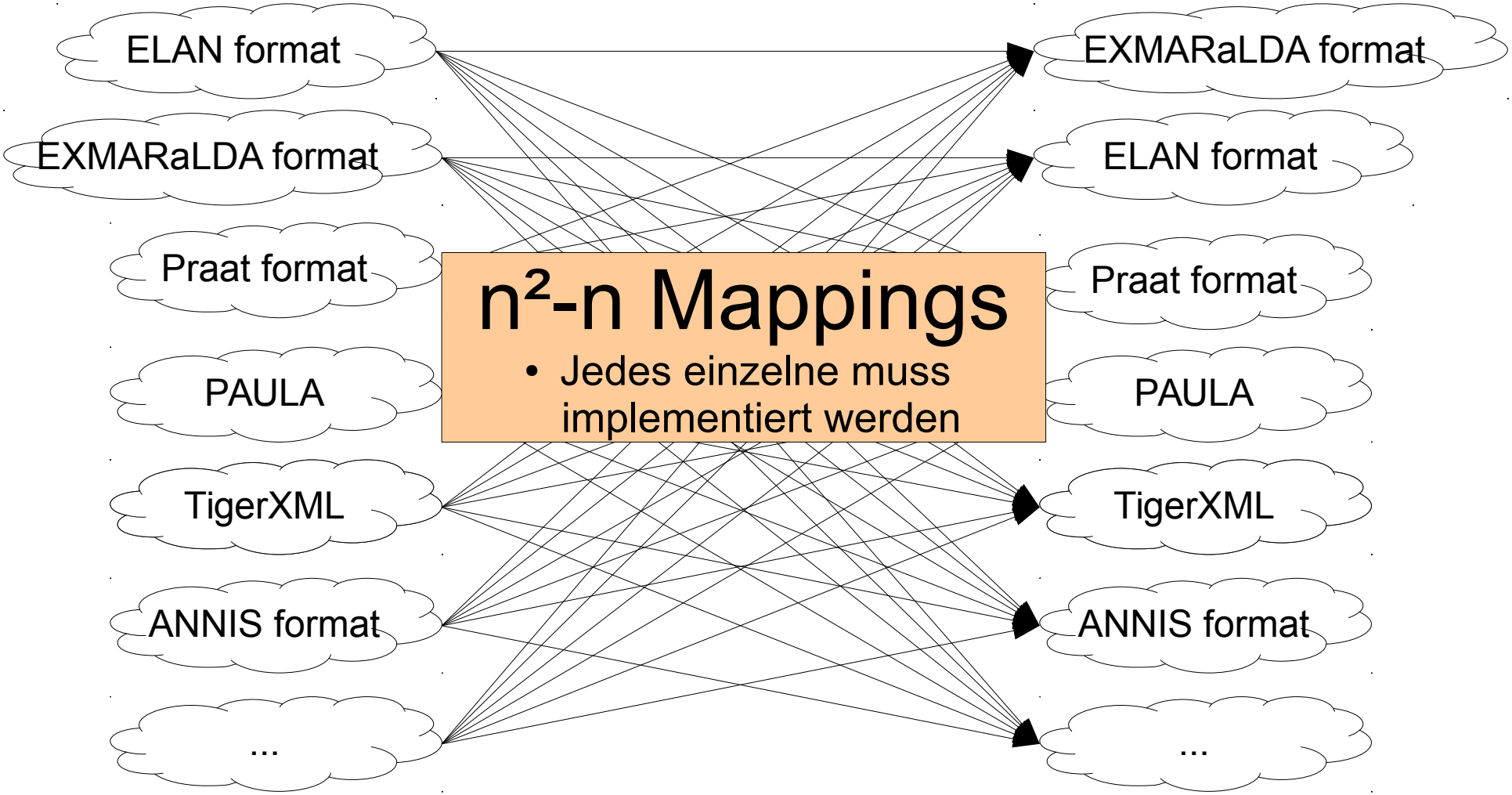
SaltNPepper

- Austausch:



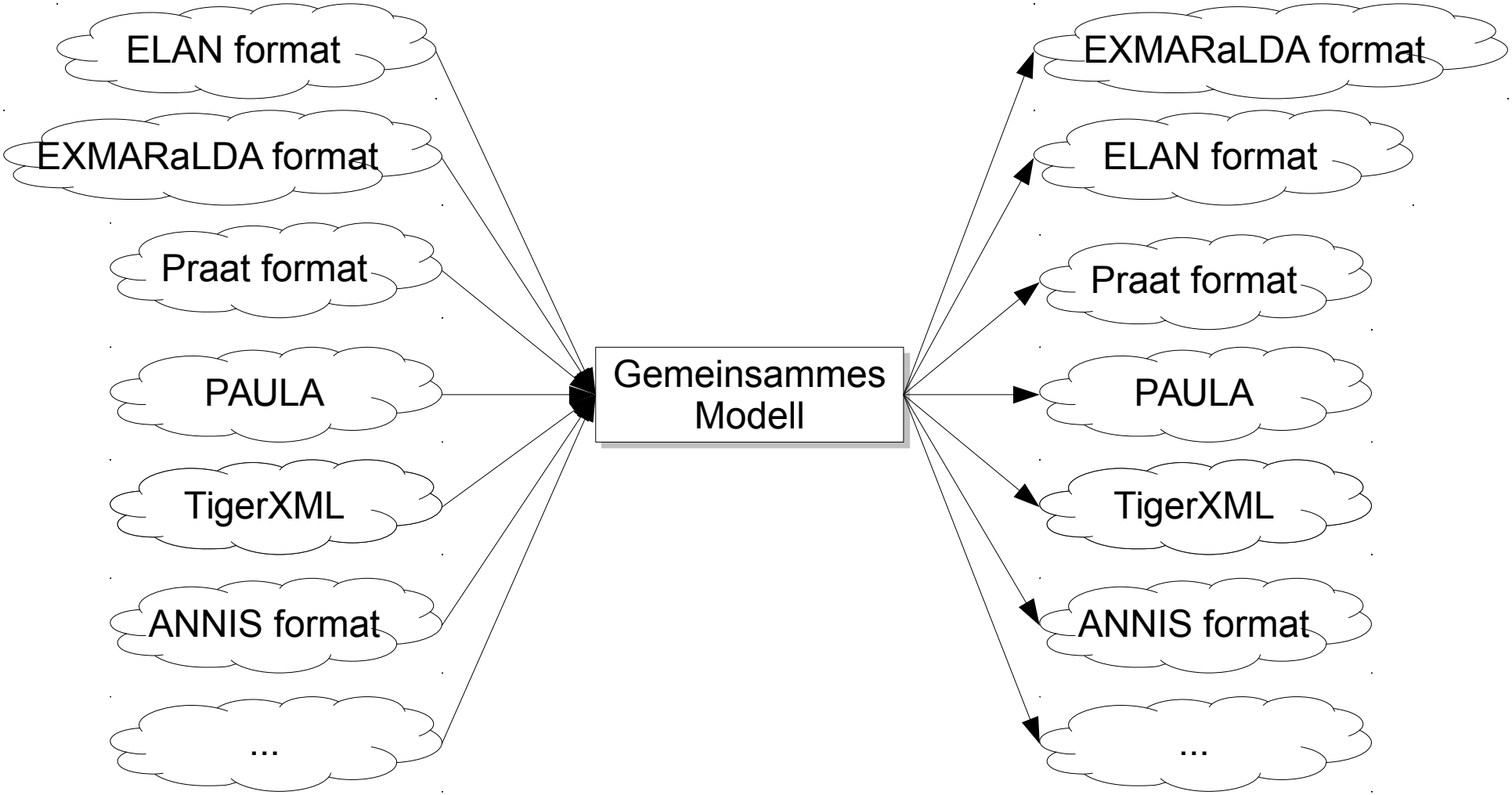
SaltNPepper

- Austausch:



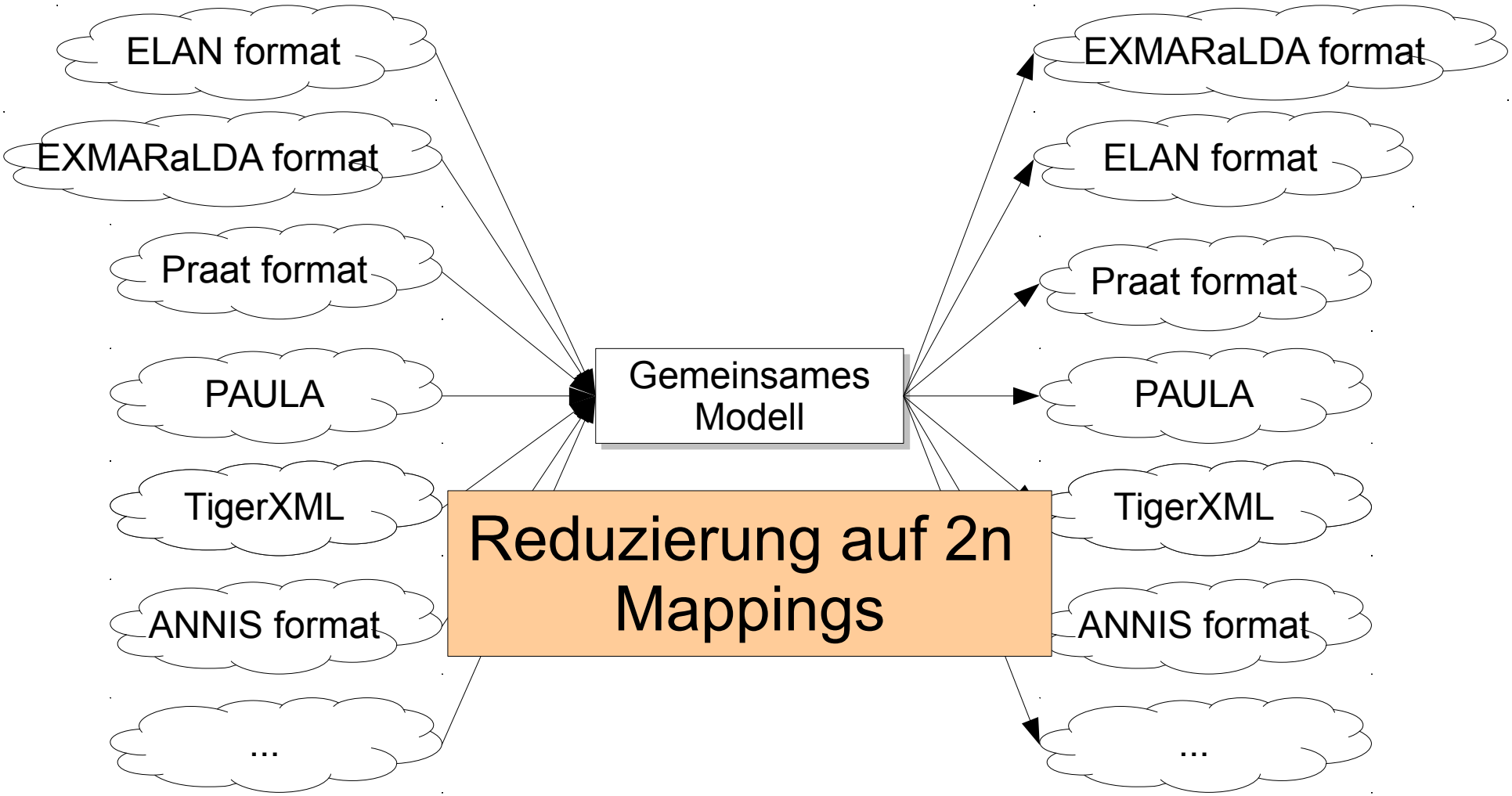
SaltNPepper

- Austausch:



SaltNPepper

- Austausch:



Salt

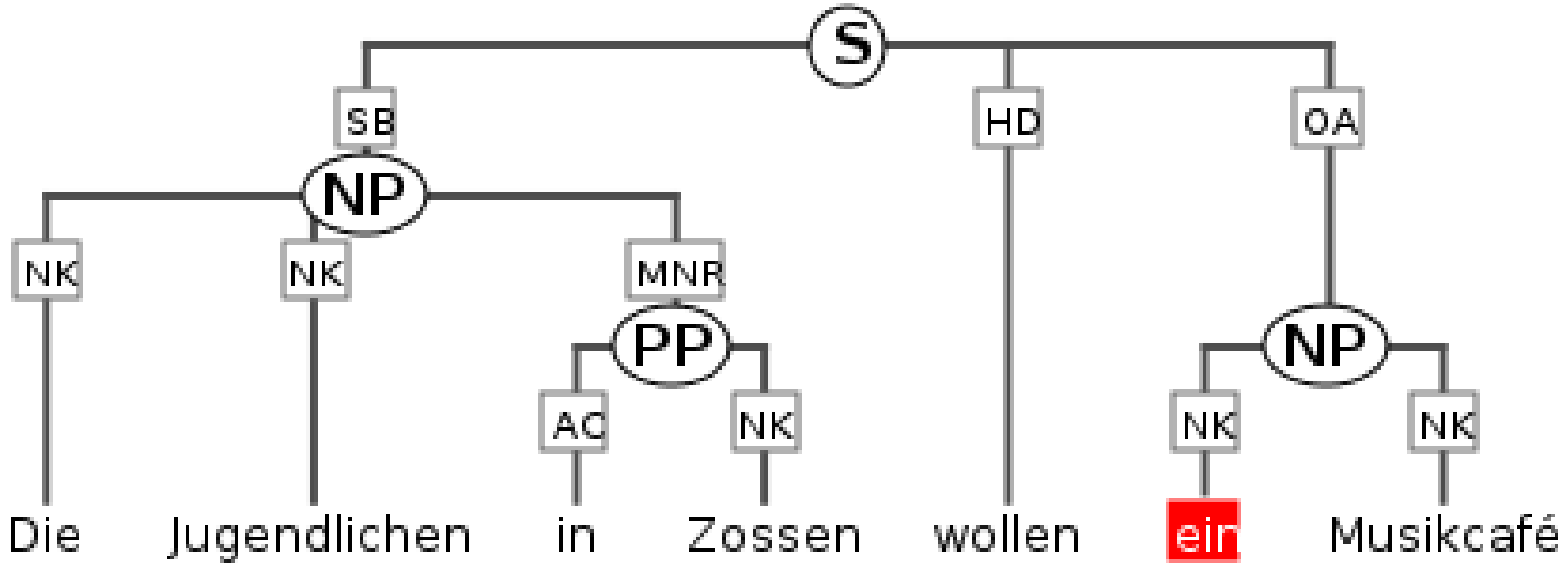
- Anforderungen an Metamodell:
  - Tagsetunabhängig
  - Beliebige Annotationsebenen
  - Unterschiedliche Annotationsarten
  - Theorieneutral

Salt

- Salt ist ein Graph?
  - Ein Graph G= (V, E) mit:
    - Einer Menge an Knoten V
    - Einer Menge an Kanten E mit e= ($v_1 \in V$, $v_2 \in V$) $\in$ E.

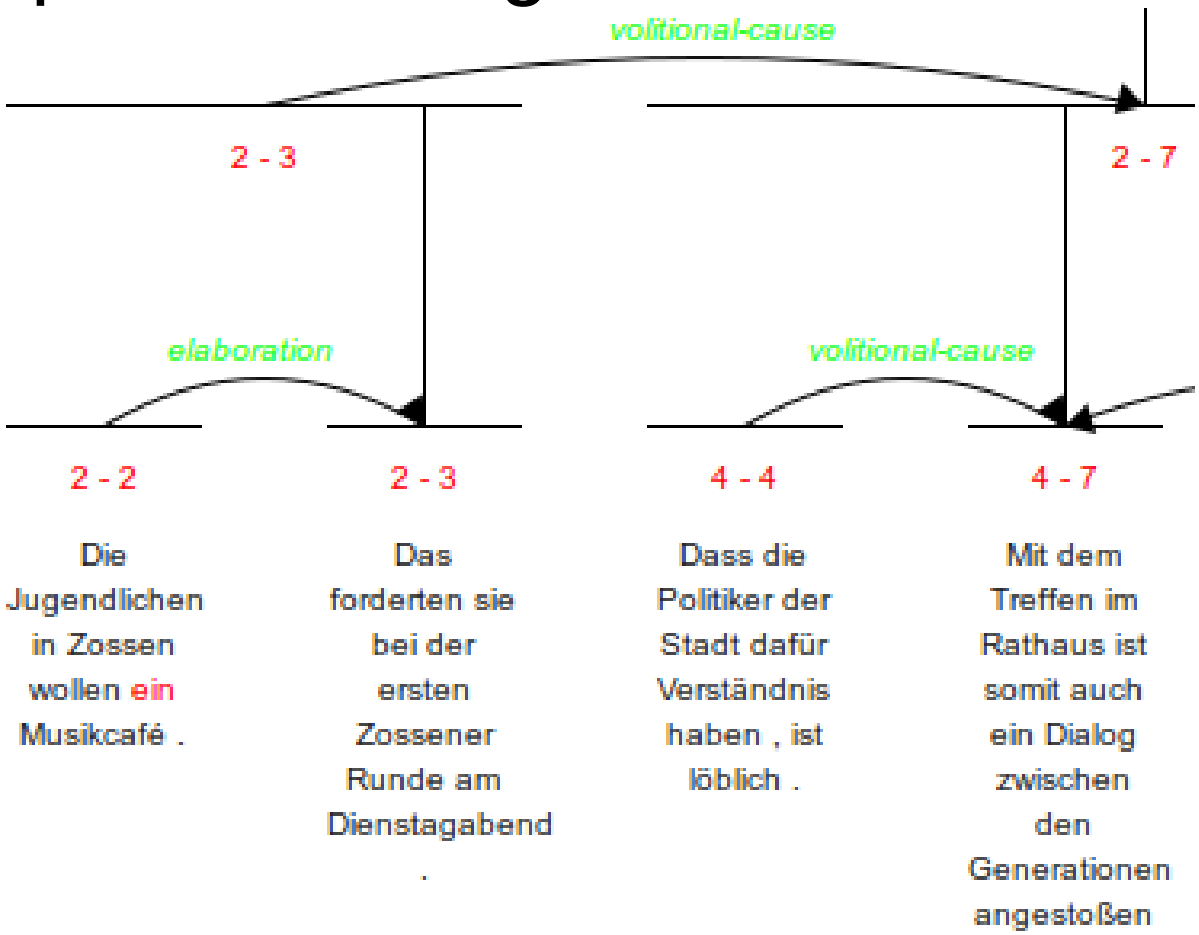
Salt

- Ein Graph in der Linguistik?



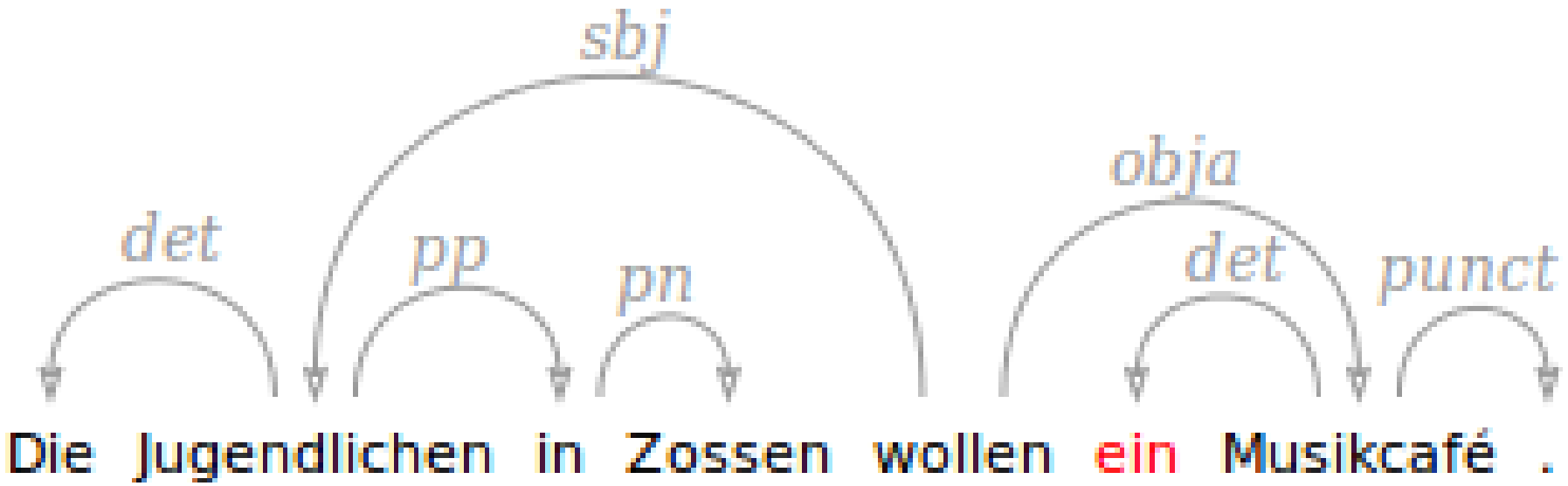
Salt

- Ein Graph in der Linguistik?



Salt

- Ein Graph in der Linguistik?



Salt

- Ist das noch ein Graph?

Feigenblatt Die Jugendlichen in Zossen wollen ein Musikcafé . Das forderten sie bei der ersten Zossener Runde am Dienstagabend . Dass die Politiker der Stadt dafür

Salt

- Und das?

Focus_newInf					nf-unsol						
Inf-Stat	new			new		new			giv-active		giv-active
NP	NP			NP		NP			NP		NP
PP			PP								
Sent	s								s		
Topic	ab										ab
tok	Die	Jugendlichen	in	Zossen	wollen	ein	Musikcafé	.	Das	forderten	sie

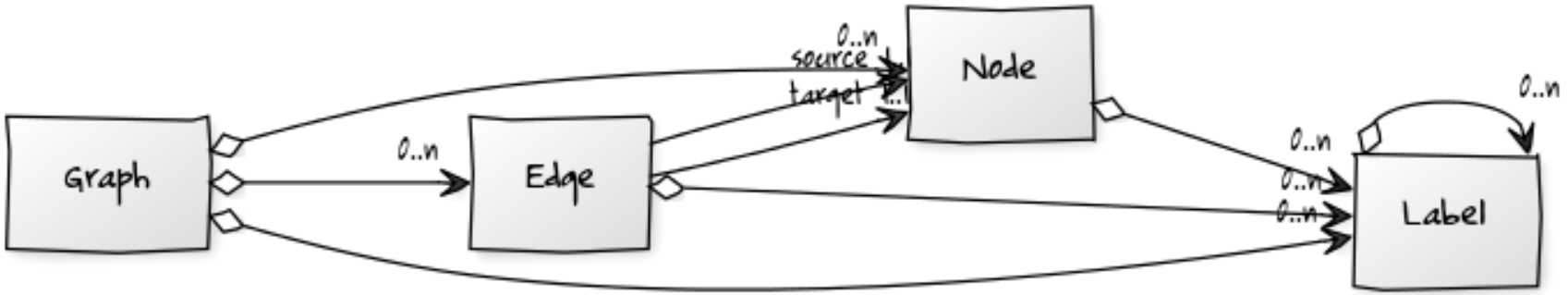
Salt

- Oder das?

Die	Jugendlichen	in	Zossen	wollen	ein	Musikcafé	.
der	jugendliche	in	Zossen	wollen	ein	Musikcafé	.
Nom.Pl.*	Nom.Pl.*	--	Dat.Sg.Neut	3.Pl.Pres.Ind	Acc.Sg.Neut	Acc.Sg.Neut	--
ART	NN	APPR	NE	VMFIN	ART	NN	$.

Salt

- Für Salt ja!



Salt

Die	Jugendlichen	in	Zossen	wollen	ein	Musikcafé	.
der	jugendliche	in	Zossen	wollen	ein	Musikcafé	.
Nom.Pl.*	Nom.Pl.*	–	Dat.Sg.Neut	3.Pl.Pres.Ind	Acc.Sg.Neut	Acc.Sg.Neut	–
ART	NN	APPR	NE	VMFIN	ART	NN	$.

Primärtext: Die Jugendlichen in Zossen wollen ein Musikcafé.

Salt

Die	Jugendlichen	in	Zossen	wollen	ein	Musikcafé	.
der	jugendliche	in	Zossen	wollen	ein	Musikcafé	.
Nom.Pl.*	Nom.Pl.*	–	Dat.Sg.Neut	3.Pl.Pres.Ind	Acc.Sg.Neut	Acc.Sg.Neut	–
ART	NN	APPR	NE	VMFIN	ART	NN	$.

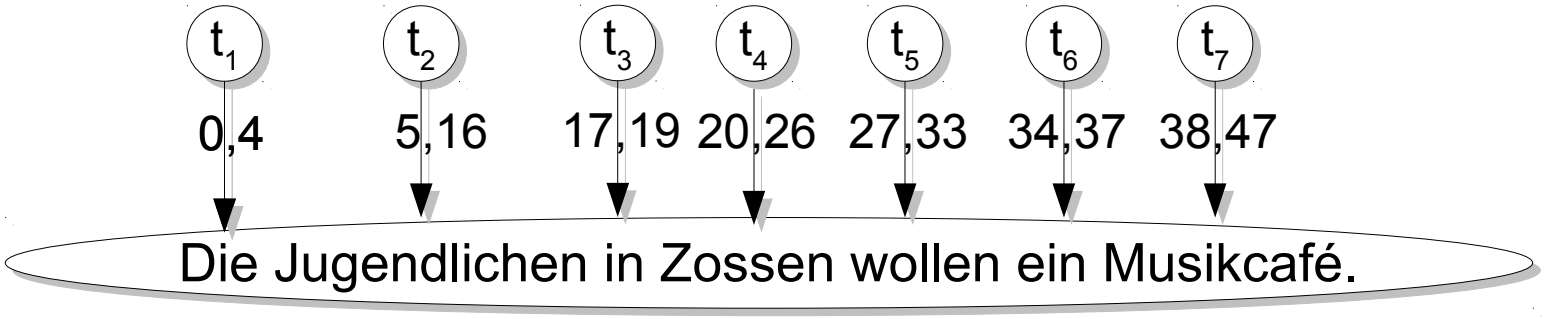
Primärtext: Die Jugendlichen in Zossen wollen ein Musikcafé.

Salt

Die	Jugendlichen	in	Zossen	wollen	ein	Musikcafé	.
der	jugendliche	in	Zossen	wollen	ein	Musikcafé	.
Nom.Pl.*	Nom.Pl.*	–	Dat.Sg.Neut	3.Pl.Pres.Ind	Acc.Sg.Neut	Acc.Sg.Neut	–
ART	NN	APPR	NE	VMFIN	ART	NN	$.

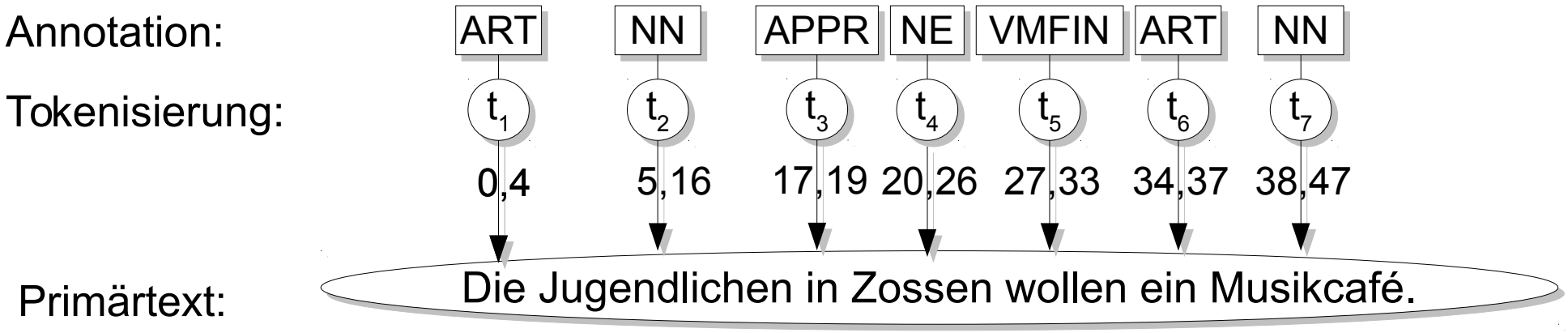
Tokenisierung:

Primärtext:



Salt

Die	Jugendlichen	in	Zossen	wollen	ein	Musikcafé	.
der	jugendliche	in	Zossen	wollen	ein	Musikcafé	.
Nom.Pl.*	Nom.Pl.*	–	Dat.Sg.Neut	3.Pl.Pres.Ind	Acc.Sg.Neut	Acc.Sg.Neut	–
ART	NN	APPR	NE	VMFIN	ART	NN	$.



Salt

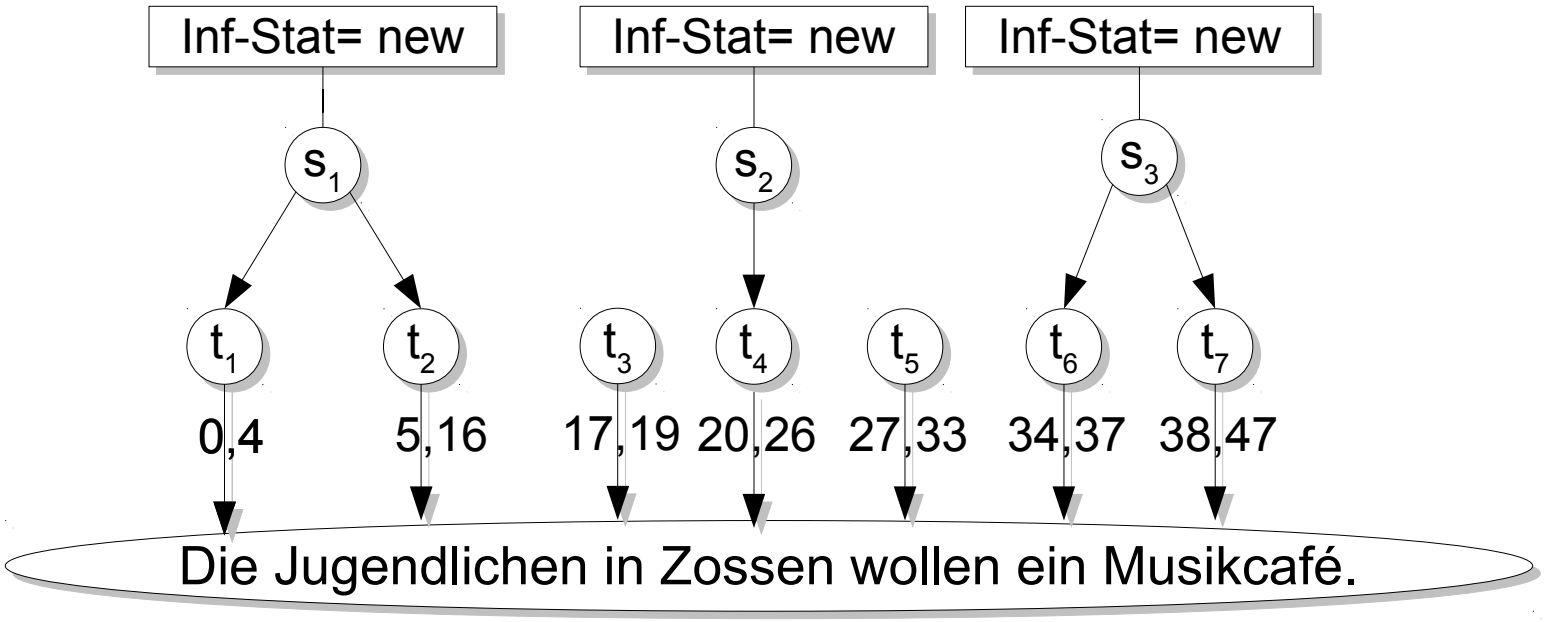
Inf-Stat	new			new			new			giv-active			giv-active
NP	NP			NP			NP			NP			NP
PP			PP										
Sent	s									s			
Topic	ab												ab
tok	Die	Jugendlichen	in	Zossen	wollen	ein	Musikcafé	.	Das	forderten	sie		

Annotation:

Mengen:

Tokenisierung:

Primärtext:

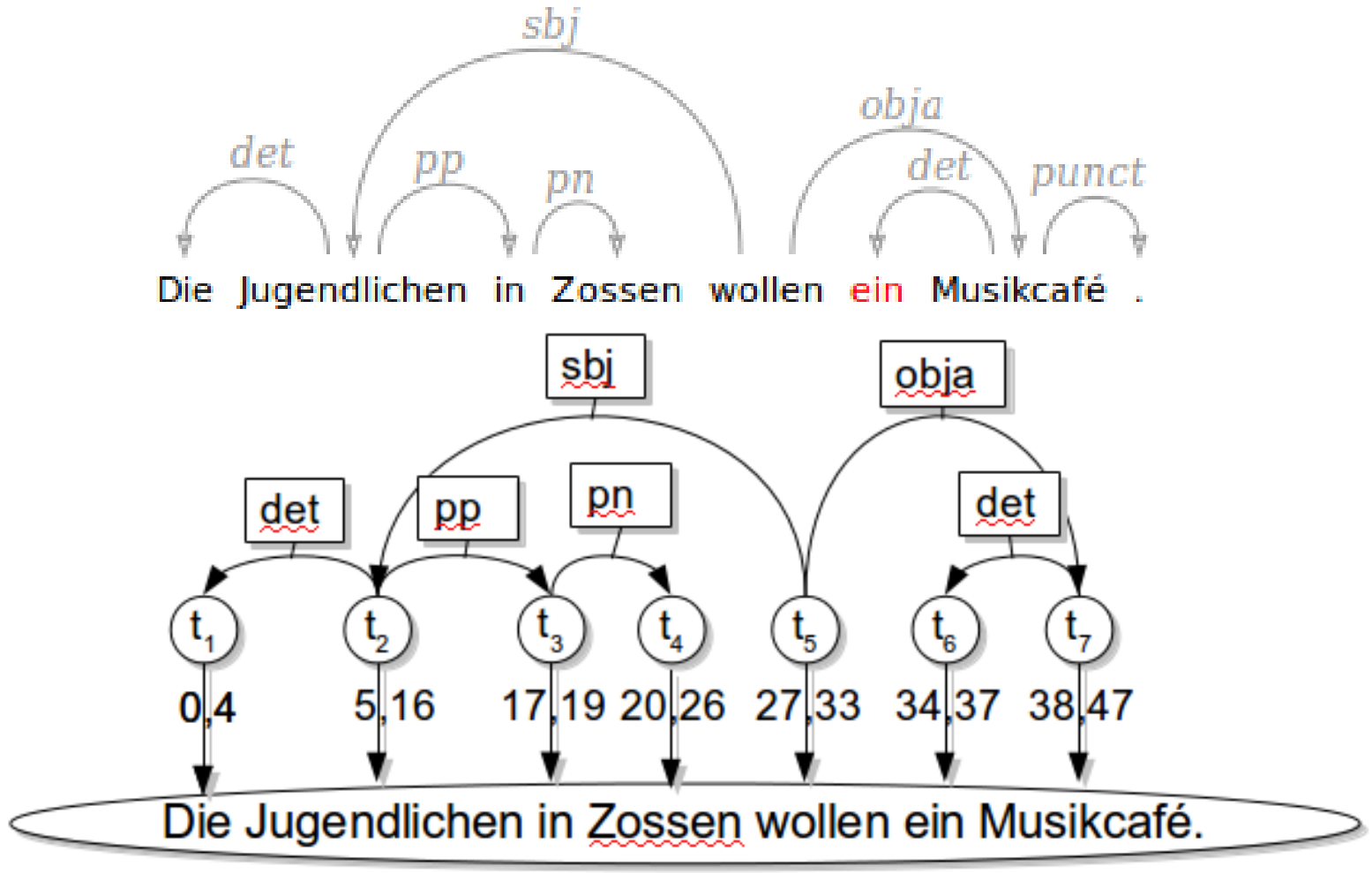


Salt

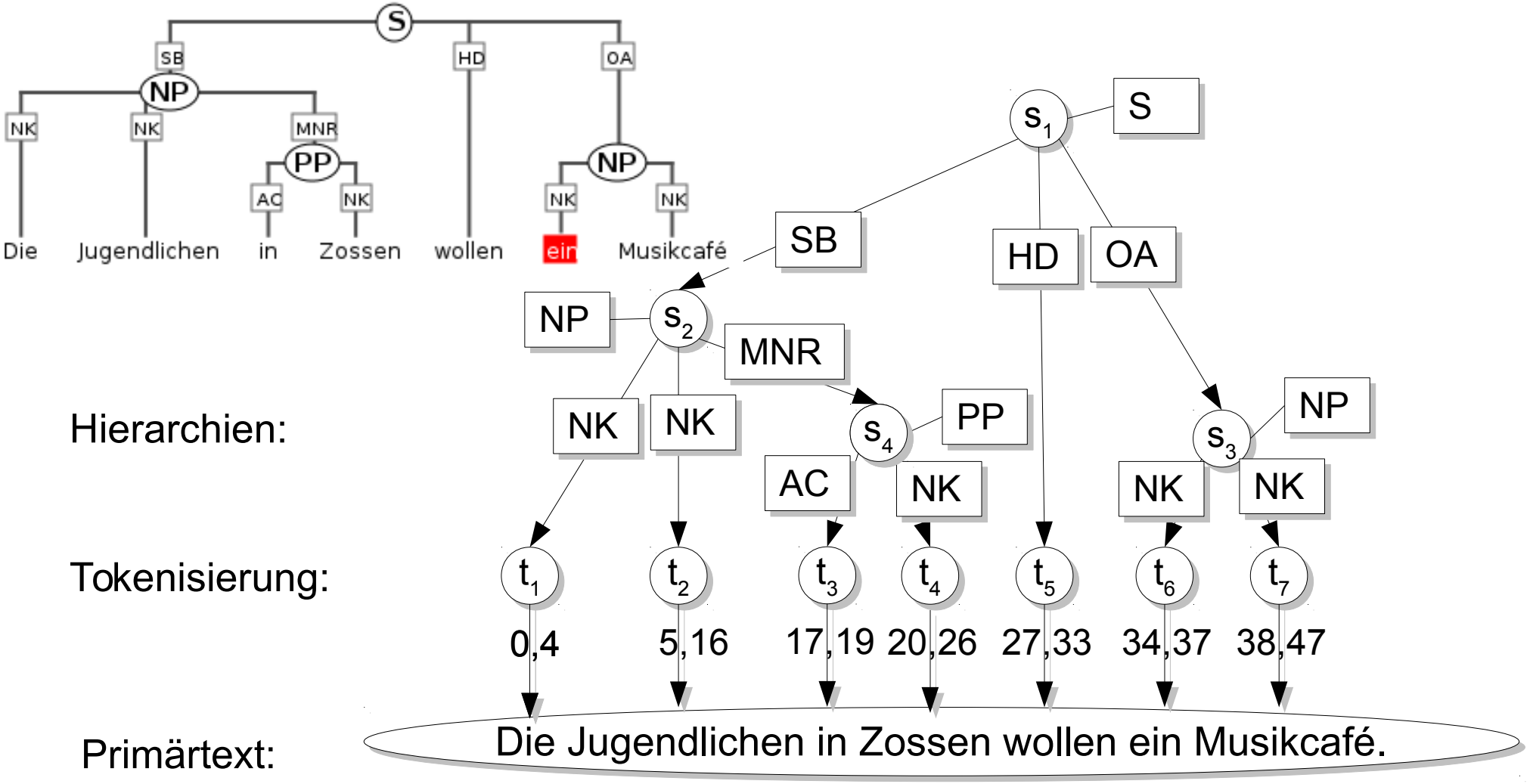
Kanten:

Tokenisierung:

Primärtext:



Salt


S
SB
HD
OA
NP
NK
NK
MNR
PP
NK
NP
AC
NK
NK
NK
Die
Jugendlichen
in
Zossen
wollen
ein
Musikcafé
s1
s2
s3
s4
t1
t2
t3
t4
t5
t6
t7
0,4
5,16
17,19
20,26
27,33
34,37
38,47
Hierarchien:
Tokenisierung:
Primärtext:
Die Jugendlichen in Zossen wollen ein Musikcafé.

Salt

- Anforderungen an Metamodell:
  - ☑ Tagsetunabhängig

    frei wählbare Attribut-Wert-Paare für Labels
  - ☑ Beliebige Annotationsebenen

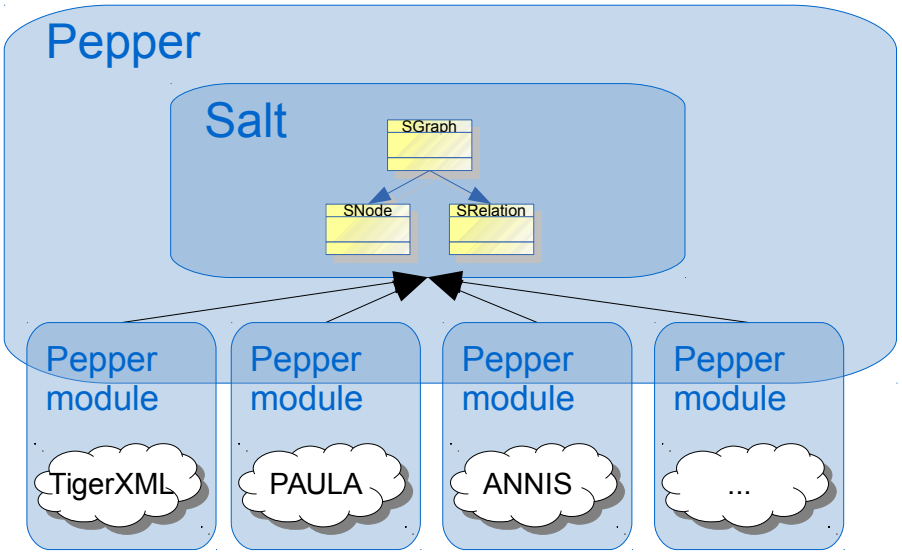
    unbegrenzte Anzahl an Labels
  - ☑ Unterschiedliche Annotationsarten

    alles, was als Graph darstellbar ist
  - ☑ Theorieneutral

    Semantikarmut, Salt kennt nur Zeichenketten

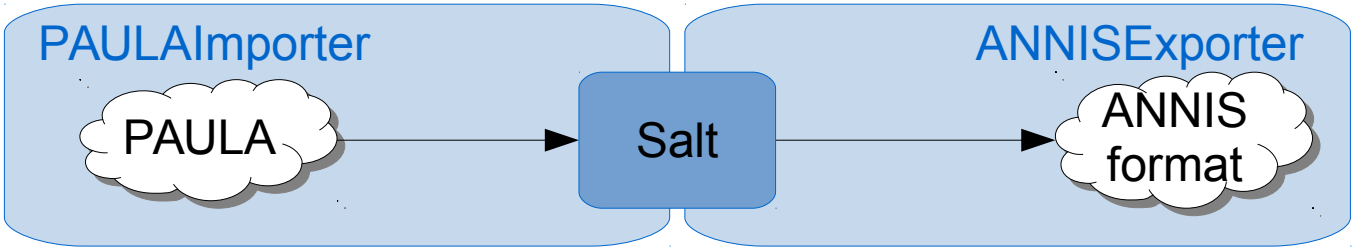
Pepper

- Pepper
  - Converterframework
  - Basiert auf Salt
  - Nur eine Infrastruktur, die Arbeit machen die Plugins



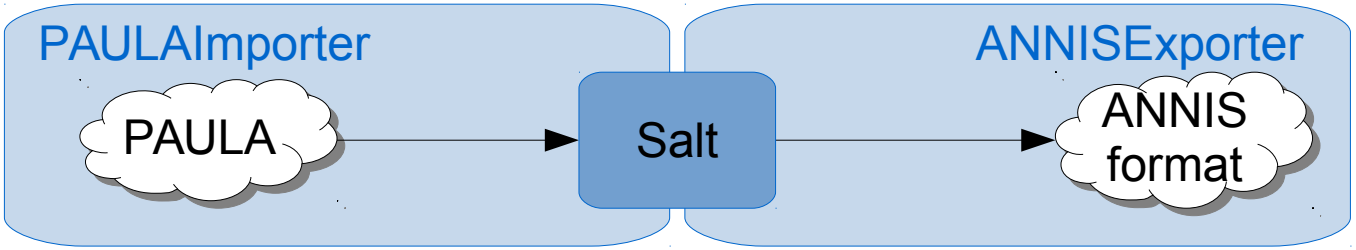
Pepper

- Drei Arten von Modulen:
  - Importer: Format A → Salt
  - Manipulator: Salt → Salt
  - Exporter: Salt → Format B

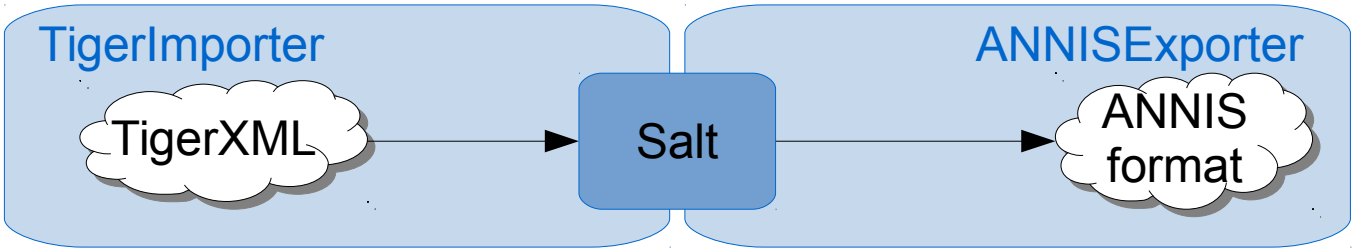


Pepper

- Drei Arten von Modulen:
  - Importer: Format A → Salt
  - Manipulator: Salt → Salt
  - Exporter: Salt → Format B

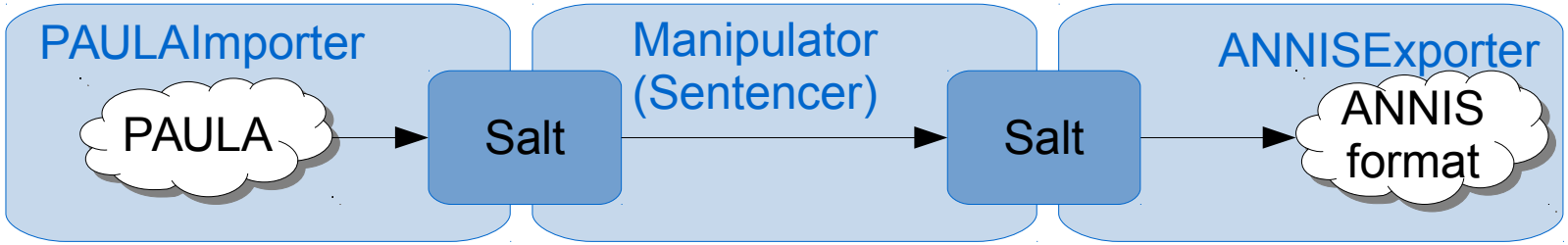


  - Kombinierbarkeit



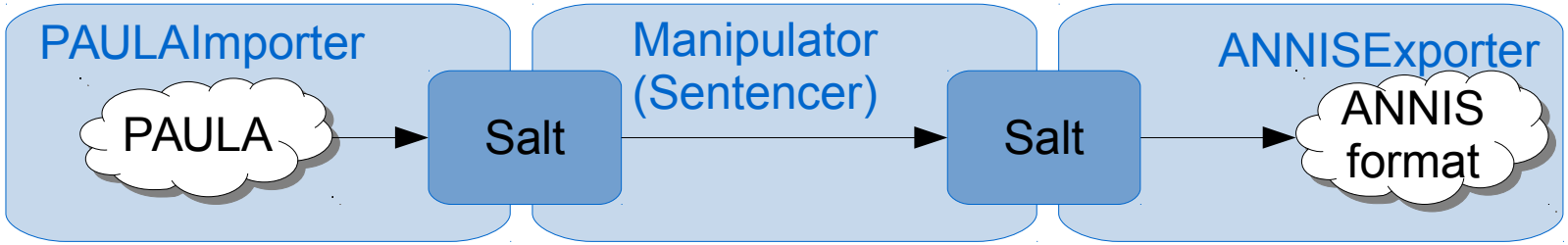
Pepper

- Manipulation

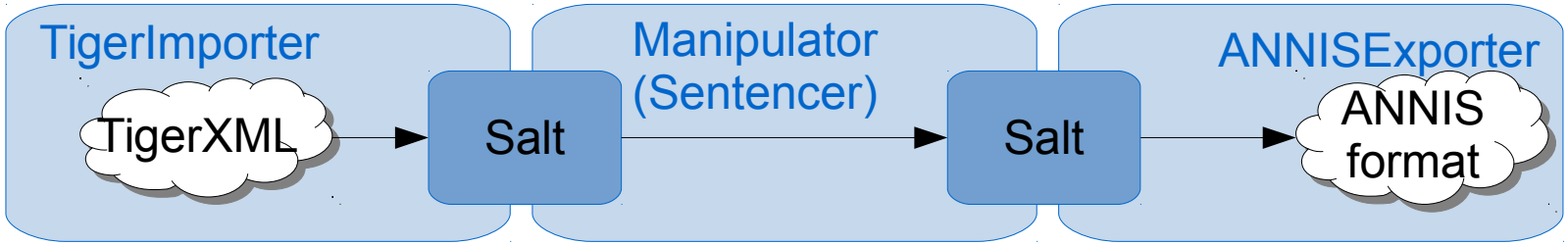


Pepper

- Manipulation



- Kombinierbarkeit



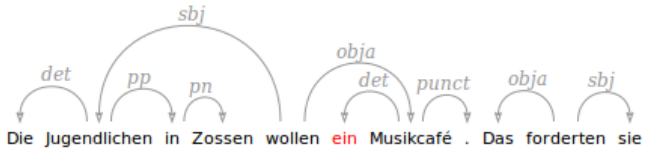
Anforderungen

- Was wir brauchen:
  - ☑ Übertragung alter Daten in neue Formate/ Standards (Nachhaltigkeit)
  - ☑ Austausch der Daten zwischen unterschiedlichen Tools (Interoperabilität)
  - ☐ Verschmelzen verschiedener Annotationsarten und -ebenen (Mehrebenenkorpora)

Mehrebenenkorpora

- Mehrebenenkorpora:

  verschiedene Annotationsarten gemeinsam durchsuchen und anzeigen

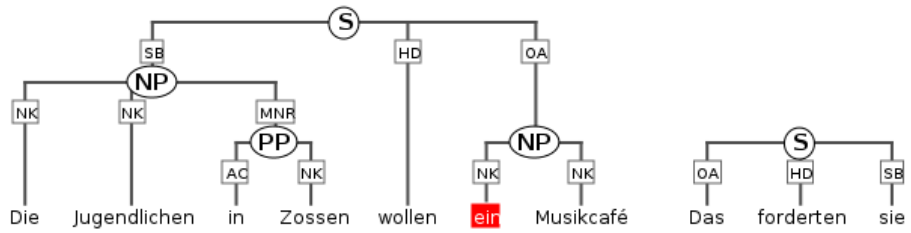


information structure (grid)

Focus_newInf						nf-unsol					
Inf-Stat	new			new		new		giv-active			giv-active
NP	NP			NP		NP		NP			NP
PP			PP								
Sent	s							s			
Topic	ab										ab
tok	Die	Jugendlichen	in	Zossen	wollen	ein	Musikcafé	.	Das	forderten	sie

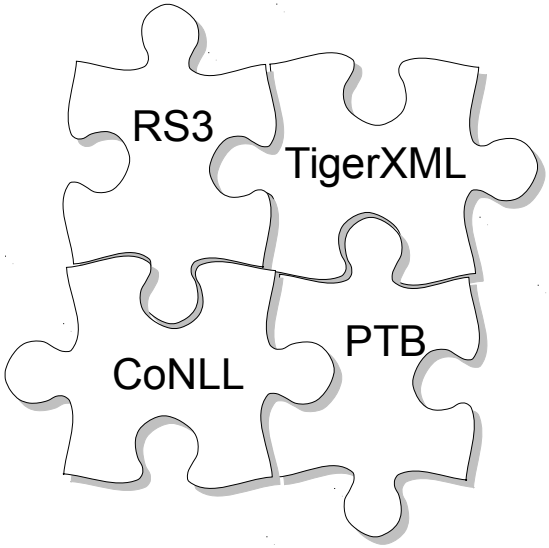
discourse referents (grid)

constituents (tree)



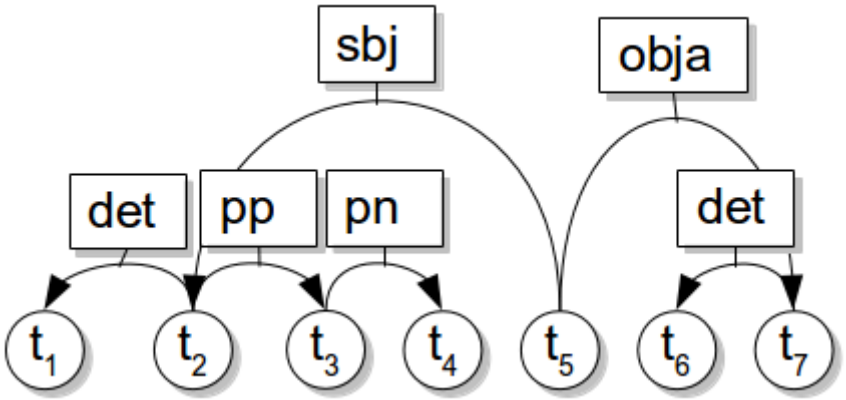
Mehrebenenkorpora

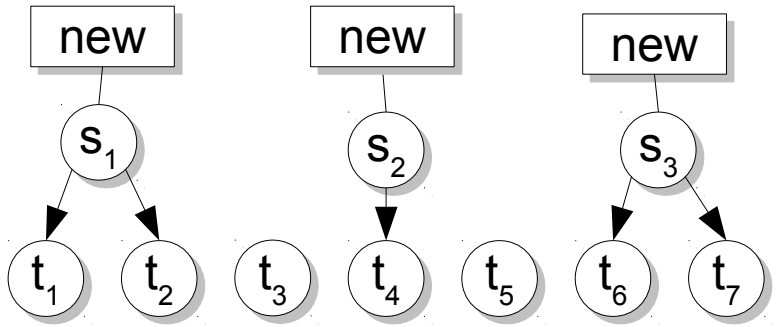
- Problem: es gibt nur wenige Mehrebenenannotationstools (bspw. WebAnno, ATOMIC)
- Idee: Verschmelzen der unterschiedlichen Formate (und somit der Ebenen)



Mehrebenenkorpora

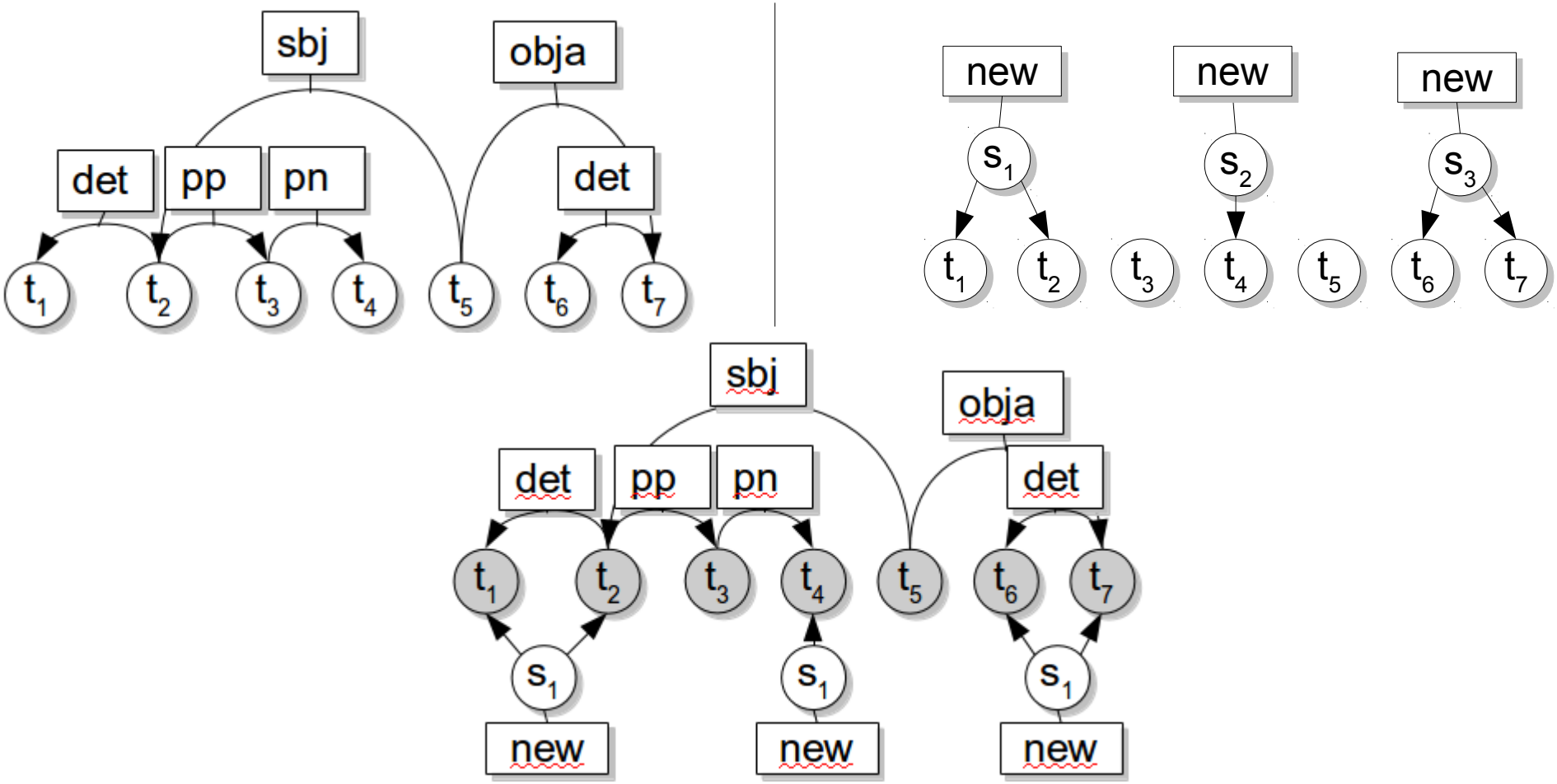
- Salt reduziert Merging zu Graphmerging





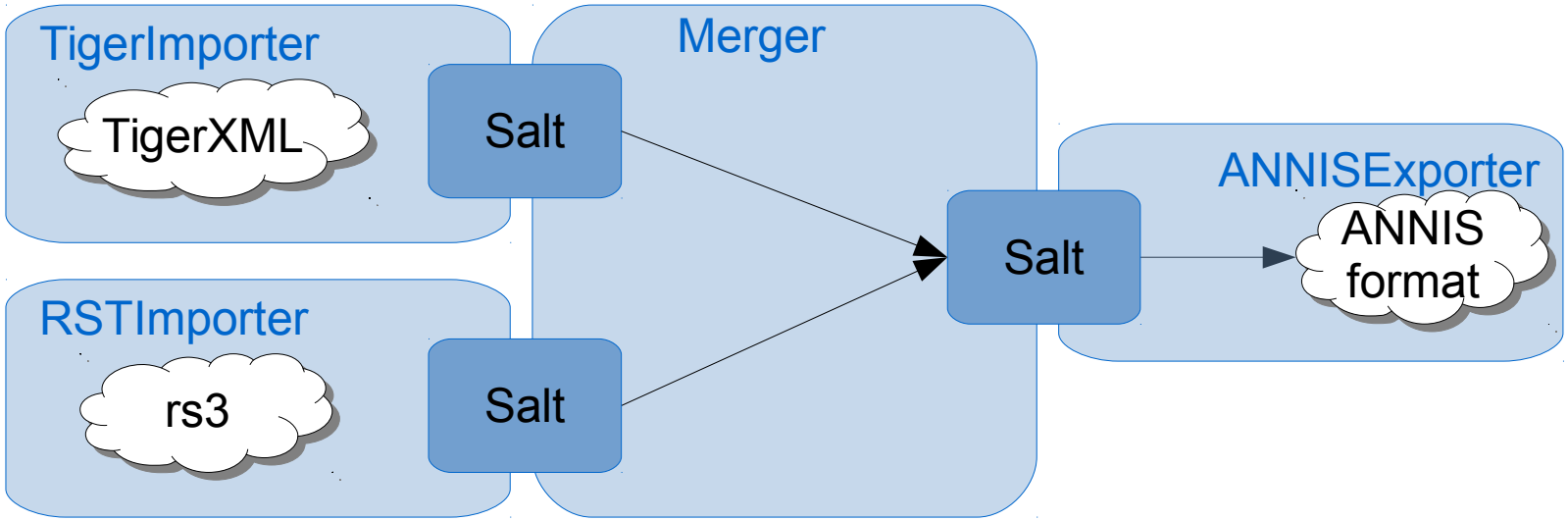
Mehrebenenkorpora

- Salt reduziert Merging zu Graphmerging



Mehrebenenkorpora

- Merger ist Plugin für Pepper (Manipulator)



Anforderungen

- Was wir brauchen:
  - ☑ Übertragung alter Daten in neue Formate/ Standards (Nachhaltigkeit)
  - ☑ Austausch der Daten zwischen unterschiedlichen Tools (Interoperabilität)
  - ☑ Verschmelzen verschiedener Annotationsarten und -ebenen (Mehrebenenkorpora)

Fazit

- SaltNPepper
  - Konvertierung von Korpora zwischen Formaten
  - Erweiterbarkeit um neue Formate (Plugins)
  - Open Source (Apache License 2.0)
  - Öffentliche Plattform: GitHub
    - https://github.com/korpling/pepper
    - https://github.com/korpling/salt

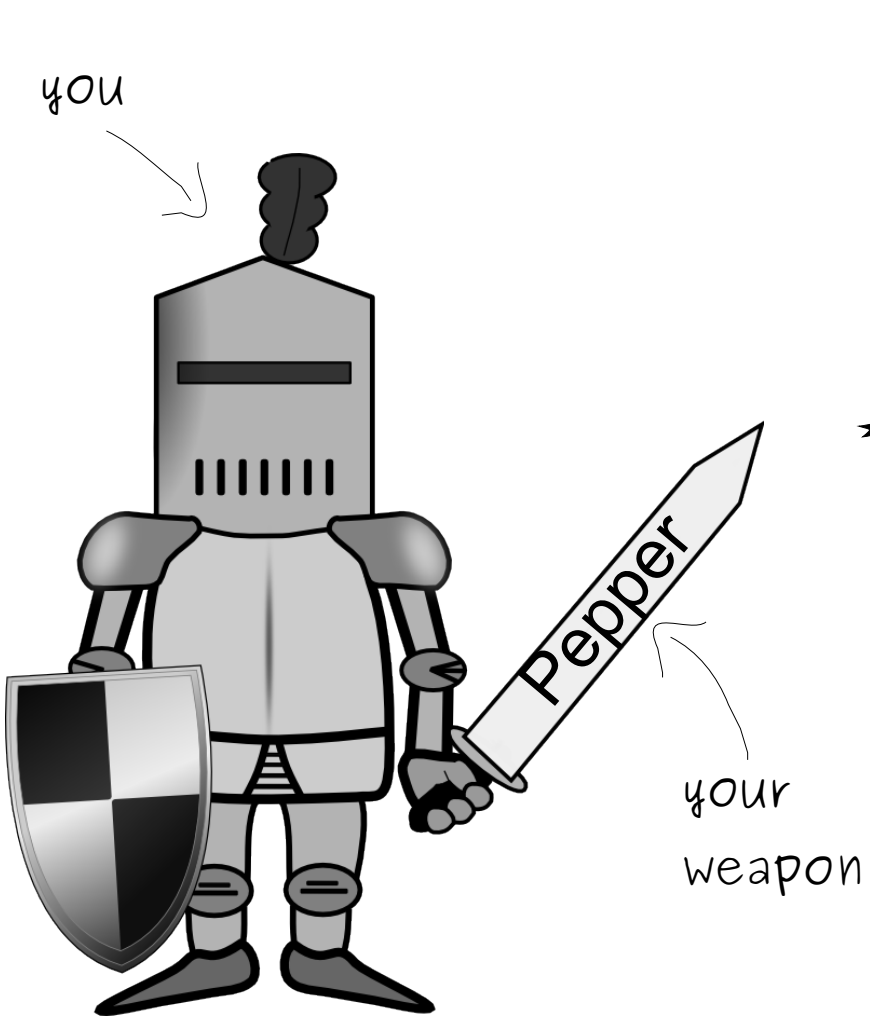
Fazit

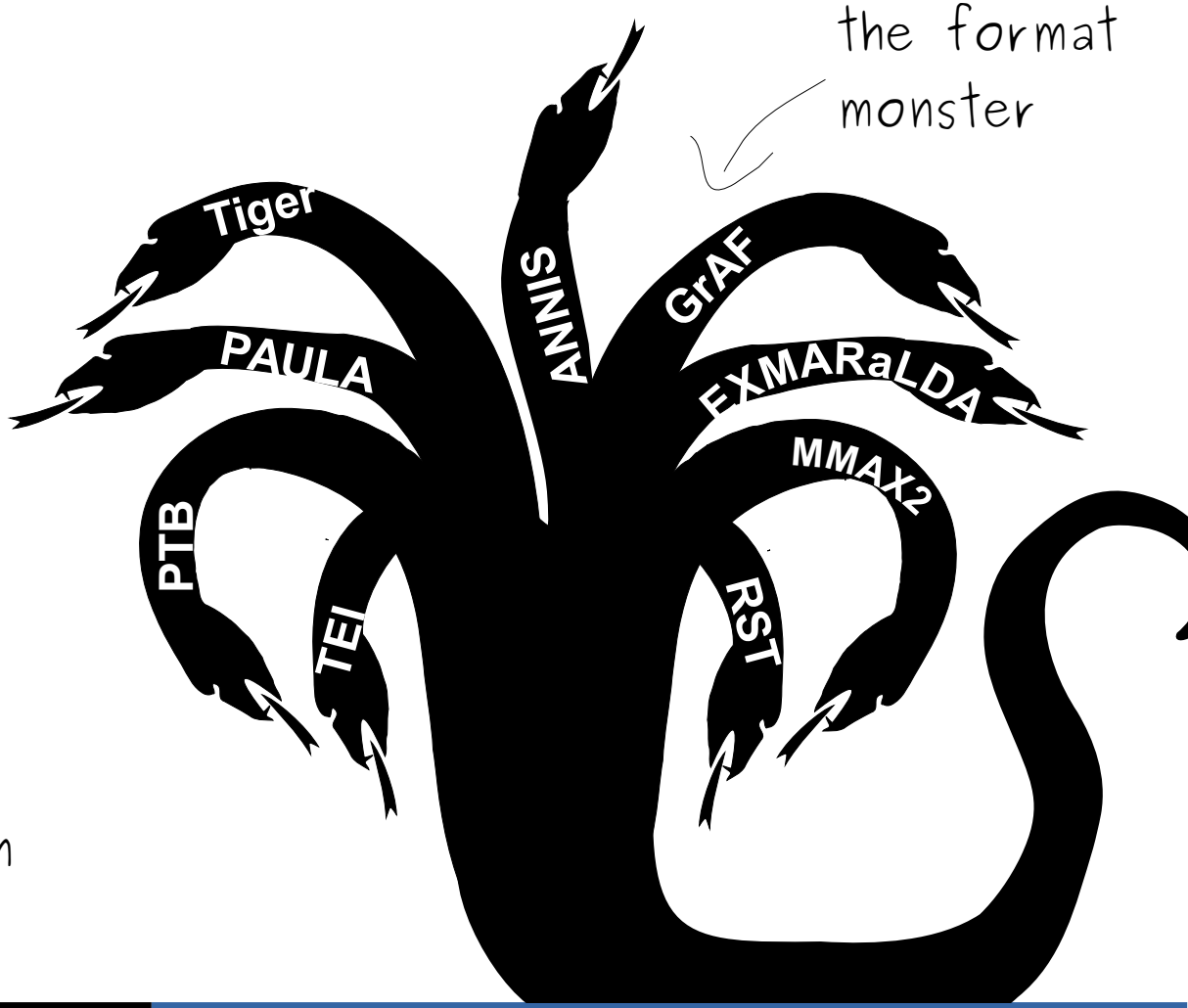
- Nachhaltigkeit von Korpora, Formaten und Software hängt zusammen
- Problem: Projekte sind befristet!
  - Oft stirbt Software nach Ende eines Projektes → Verlust von Geld und Zeit

Fazit

- Nachhaltigkeit von Korpora, Formaten und Software hängt zusammen
- Problem: Projekte sind befristet!
  - Oft stirbt Software nach Ende eines Projektes → Verlust von Geld und Zeit
- Software braucht zum Überleben:
  - Aktive Entwicklercommunity
  - Open Source
  - Öffentliche Plattform
  - Gute Dokumentation

Danke für Ihre Aufmerksamkeit





SaltNPepper und das Formatpluriversum

- Diese Folien wurden erstellt unter Verwendung von:
  - Yuml http://yuml.me
  - Openclipart http://openclipart.org