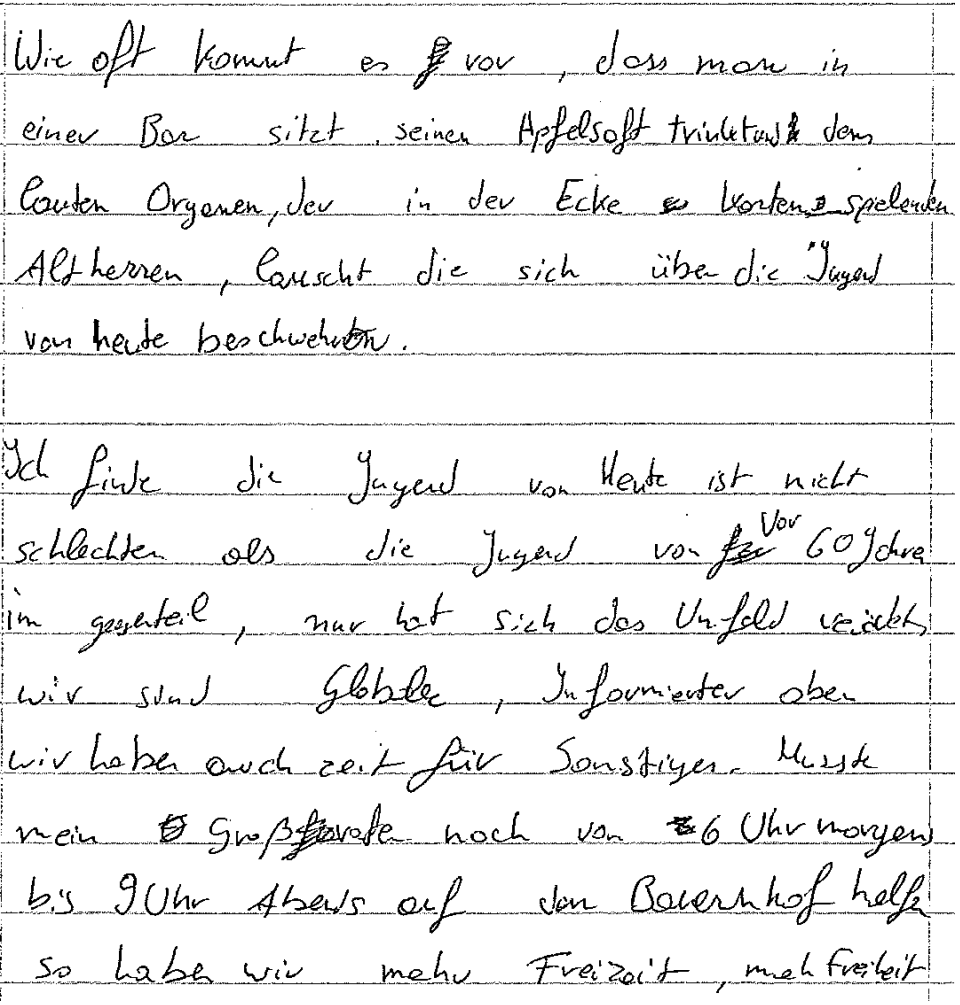
Automatische Annotation von Schülertexten

Herausforderungen und Lösungsvorschläge am Beispiel des Projekts KoKo

Andrea Abel, Aivars Glaznieks, Egon W. Stemle

Institut für Fachkommunikation und Mehrsprachigkeit,
Europäische Akademie Bozen (EURAC)



Wie oft kommt es f vor, dass man in einer Bar sitzt, seinen Apfelsaft trinkt und l *dem* lauten Organen, der in der Ecke s kartenspielenden Altherren, lauscht, die sich über die Jugend von heute beschwehren. <ABSATZ>

<LEERZEILE>

Ich finde, die Jugend von Heute ist nicht schlechter als die Jugend von for vor 60 Jahren, im gegenteil, nur hat sich das Umfeld verändert, wir sind Globaler, Informierter, aber wir haben auch zeit für Sonstiges. Musste mein O Großfavater noch von ? 6 Uhr morgens bis 9 Uhr Abends auf *den* Bauernhof helfen, so haben wir mehr Freizeit, meh Freiheit

Hintergrund I

Merkmale:

- Fehler: Orthographie, Interpunktion, Morphosyntax, Lexik
- „Auffälligkeiten“: Lexik (v.a. formelhafte Sprache), Morphosyntax (Mündlichkeit), Lexik (Mündlichkeit), textuelle Ebene (Kohäsion, Kohärenz)

Schülertexte als Lernertexte:

> „L1-LernerInnen sind solche, die ihre Erstsprache(n) oder wesentliche Teile davon, wie etwa Schreib- und Textkompetenzen, erlernen.“

Hintergrund II

L2-Lernerkorpora:

- Fehlerannotiert (inline), z.B. FRIDA, vgl. Granger 2003;
- Fehlerannotiert (mehrebenen, stand-off), z.B. FALKO, vgl. Lüdeling et al. 2005, Reznicek et al. i.E.; ALeSKo, vgl. Zinsmeister & Breckle i.E.; CzeSL, vgl. Hana et al. 2010, 2012)

L1-Lernertexte:

- Keine korpuslinguistisch aufbereiteten, abfragebaren Korpora für L1(DE) vorhanden
- Schülertextesammlungen (u.a. Augst et al. 2007, Thelen 2000, Berg et al. 2010, Fix & Melenk 2002, Sieber 1998): Analyse nach „Analyseraster“ (z.B. Nussbaumer & Sieber 1994)

Projekt “KoKo”

“Bildungssprache im Vergleich: korpusunterstützte Analyse der Sprachkompetenz bei Lernenden im deutschen Sprachraum (unter besonderer Berücksichtigung des Deutschen in Südtirol)”

http://www.korpus-suedtirol.it/bildungssprache_de.htm

Partner: EURAC research

Freie Universität Bozen
Libera Università di Bolzano
Free University of Bozen · Bolzano

universität innsbruck

Im Rahmen der Initiative: Korpus Südtirol
http://www.korpus-suedtirol.it

Korpus Südtirol

Ziele

1 Aussagen über **Schreibkompetenzen** von SchülerInnen

- Deutsch als Erst-/Unterrichtssprache
- Jugendliche mit höherem Bildungsniveau

2 Analyse von **Kontextvariablen**

- Diatopische und biographische Faktoren

3 Aufbau eines digitalen **Lernerkorpus** und Entwicklung von Methoden

- wissenschaftlich fundierte Basis für Erfassung und Analyse schriftlicher Sprachproduktion und nachhaltige Dokumentation
- Entwicklung von computerlinguistischen Werkzeugen zur Unterstützung der Vergleiche von (Sub-)Korpora (z.B. Anstein 2013)

Design

TeilnehmerInnen: ca. 1500 OberschülerInnen 1 Jahr vor der Matura bzw. dem Abitur

- Südtirol ca. 520
- Nordtirol ca. 460
- Thüringen ca. 520

Stratifizierte Zufallsstichprobe (nach Schulklassen) für jede Region mit folgenden Schichtungsmerkmalen:

- Stadtgröße (Schule): Groß-, Mittel- u. Kleinstadt
- Schultyp: allgemeinbildend vs. berufs-/fachspezifisch

Zeitpunkt der Datenerhebung: Stichtag 25.05.2011

Methode

- **Textproduktion** in der Schule (argumentative Texte zu einem vorgegebenen Thema)
- **Schüler- und Lehrerfragebogen (Metadatenerhebung)**
  - Kontrollvariablen (Geschlecht, Alter, Schultyp …)
  - Sozio-ökonomischer Hintergrund
  - Sprachliche Biographie der TeilnehmerInnen
- **Datenaufbereitung** und Korpuserstellung (Transkription/ Annotation, Tokenisierung, Lemmatisierung, POS-Tagging (Schmid 1994), IMS work bench (Christ 1994))
- Qualitative, **manuelle Analyse** (ling. Annotation (stand-off) nach Analyseraster) & quantitative, **halb-automatische Analyse** (korpuslinguistische Tools)
- Verknüpfung von **Textanalyse und Metadaten**

Korpuszusammensetzung

Subkorpus (nach Region)	Gesamtanzahl	L1 Deutsch
Nordtirol (NT):	233.098 Tokens (457 Texte)	206.439 Tokens (404 Texte)
Südtirol (ST):	222.209 Tokens (520 Texte)	192.891 Tokens (451 Texte)
Thüringen (TH):	353.674 Tokens (521 Texte)	317.075 Tokens (464 Texte)
ohne Angabe	2.349 Tokens (5 Texte)	---
gesamt	811.330 Tokens (1503 Texte)	716.405 Tokens 1319 (Texte)

Verfügbare Metadaten:
Muttersprache (L1=deutsch vs. Nicht deutsch), **Region** (NT vs. ST vs. TH), **Geschlecht (**weiblich vs. Männlich), **Schultyp (**AHS vs. BHS), **Klasse (**ID-Nr.), **Autor** (ID-Nr.), **Projekt, Transkribierer, Sprache**

Version „KoKo_2“: Stand Dezember 2012

Ziele beim Korpusaufbau (KoKo_2)

- Korrekte **Transkription und manuelle Annotation** der Texteigenschaften als Grundlage für die Korpuserstellung und weiterer linguistischer Annotation
- **Automatische Annotationen:**
  - Lemmata möglichst vollständig und korrekt
  - POS-Tags möglichst korrekt
  - Wiederholbarkeit der Annotation für unterschiedliche Projektphasen: möglichst keine manuellen Eingriffe
- Korpus beinhaltet alle **originalen Wortformen** (inkl. möglicher nicht standardkonformen Schreibungen)

Herausforderungen

Transkription / Annotation:

- Texteigenschaften: Selbstkorrekturen, graphische Gestaltung des Textes u.Ä.

Automatische Verarbeitung:

- Nicht-Standardschreibung
  - orthographische Fehler (v.a. Groß-/Kleinschreibung)
  - Abkürzungen (okkasionelle Abkürzungen: *monatl.*, *vllt.*, geschlechtsneutrale Schreibung: *Freund/in* u.Ä.)
  - fehlende/zusätzliche Interpunktionszeichen (u.a. Emoticons)
- Ad-hoc-Bildungen (z.B. *Klamottensucht*)
- Argumentative Texte (Erörterungen, Stellungnahmen)

Transkription + manuelle Annotation I

Annotation **während** der Transkription:

- Annotation nicht standardkonformer Schreibungen:
  - Orthographische Fehler: *error (originalForm)*
  - „Ungewöhnliche“ Abkürzungen: *reduction (reducedForm)*
- Hinzufügen einer Zielhypothese für nicht standardkonformer Schreibung:
  - *error (targetForm), reduction (unfoldedForm)*
- Hinzufügen von Zusatzinformation bei Transkriptionsunsicherheiten:
  - *ambiguous, alternative, unreadable, comment*

Bisher **keine Korrektur** der Interpunktion.

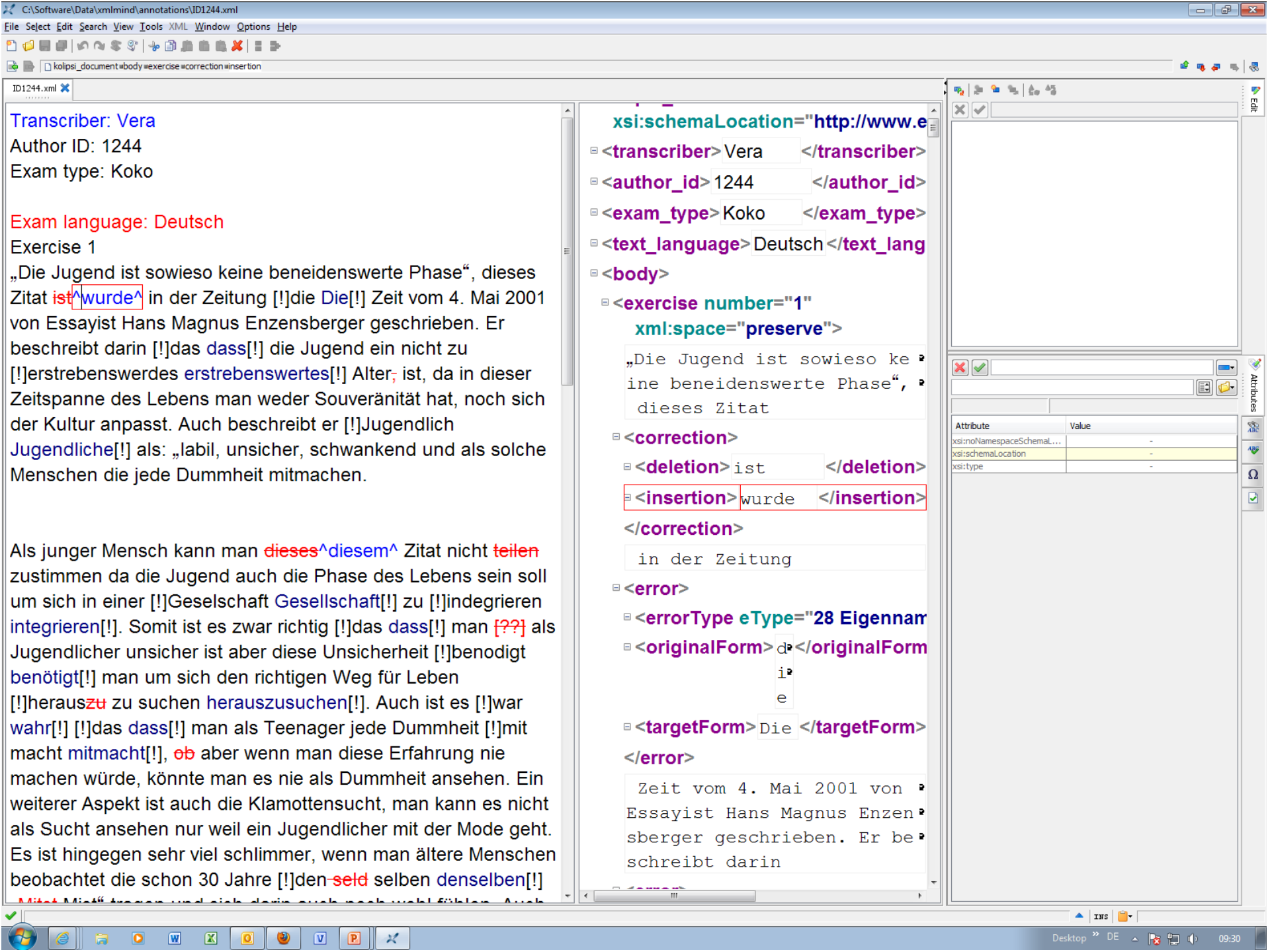
Transkription + manuelle Annotation II

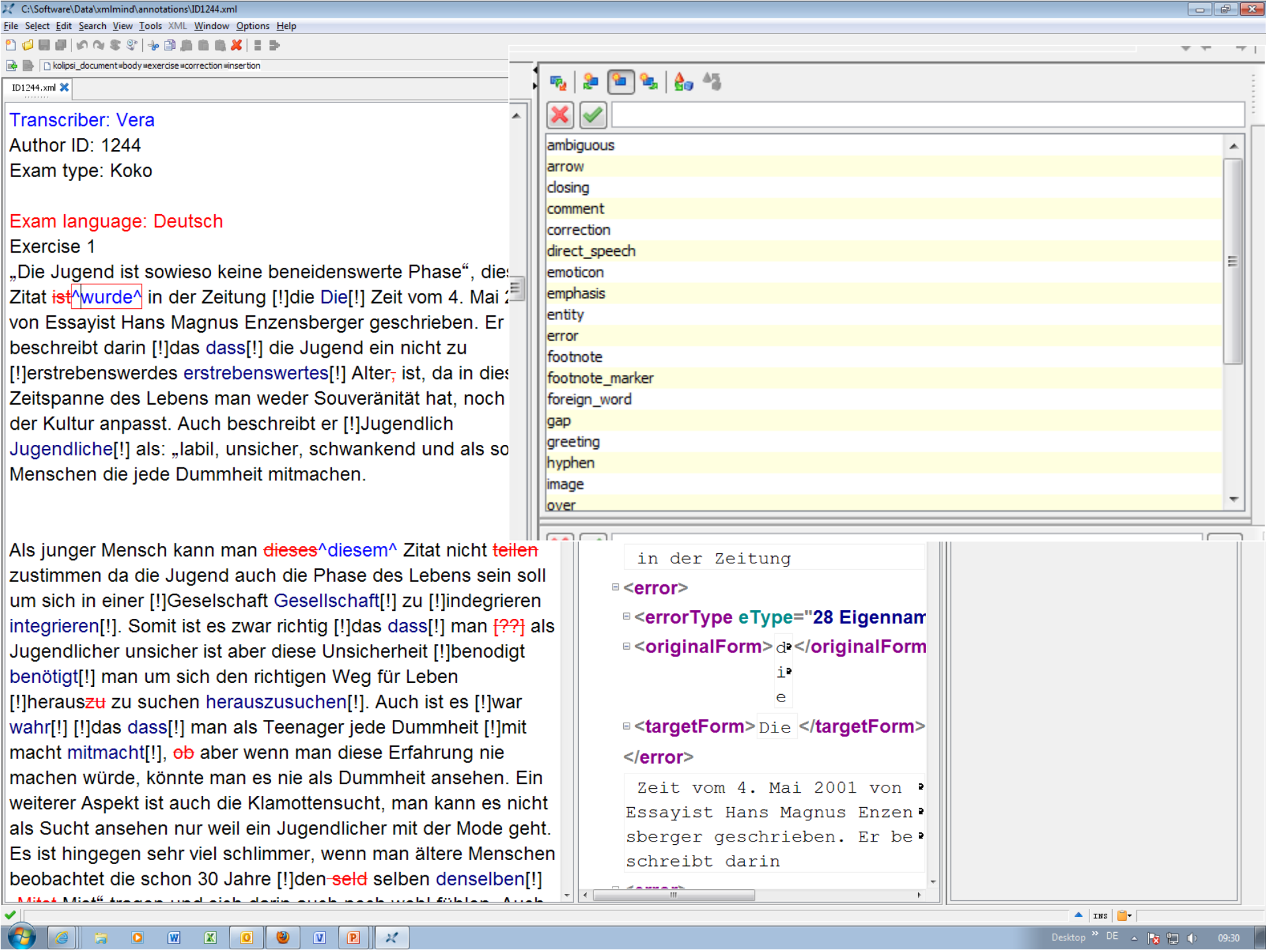
Annotation der **Texteigenschaften**:

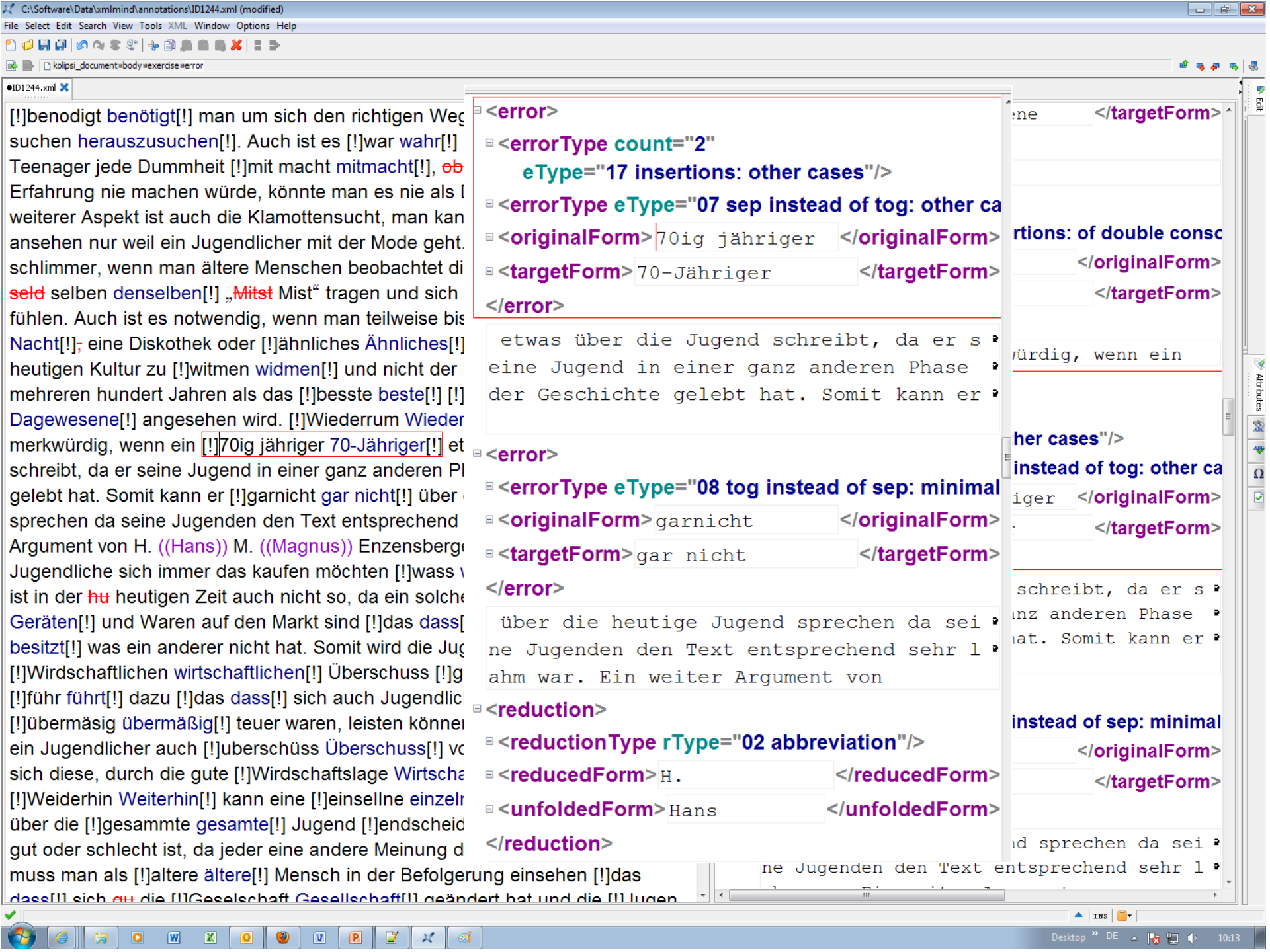
- *correction, deletion, emoticon, emphasis, error (originalForm), footnote, insertion, label, outline, paragraph, postscript, reduction (reducedForm), title, symbol, …*

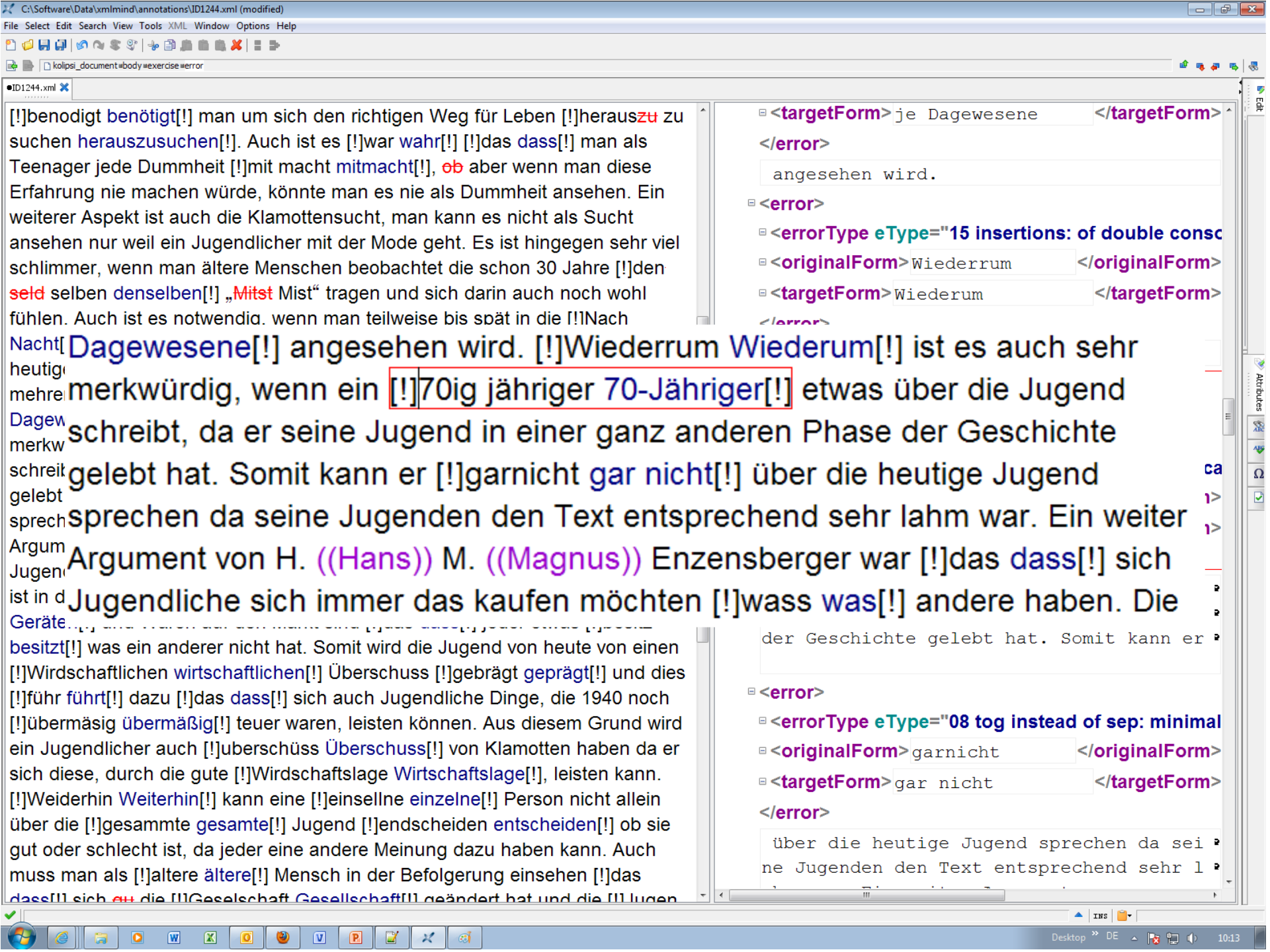
Transkription mit **XMLmind**: (vormals) frei zugänglicher XML Editor

- **Anpassung der GUI** nach den Bedürfnissen des Transkriptors / Annotators
- Projektspezifische Organisation: **Template-Dateien**
- Annotationen (Texteigenschaften) mithilfe von **Schemadateien**


C:\Software\Data\xmlmind\annotations\ID1244.xml
File Select Edit Search View Tools XML Window Options Help
kolipsi_document body exercise correction insertion
ID1244.xml
Transcriber: Vera
Author ID: 1244
Exam type: Koko
Exam language: Deutsch
Exercise 1
„Die Jugend ist sowieso keine beneidenswerte Phase“, dieses Zitat ist^wurde^ in der Zeitung [!]die Die[!] Zeit vom 4. Mai 2001 von Essayist Hans Magnus Enzensberger geschrieben. Er beschreibt darin [!]das dass[!] die Jugend ein nicht zu [!]erstrebenswerdes erstrebenswertes[!] Alter, ist, da in dieser Zeitspanne des Lebens man weder Souveränität hat, noch sich der Kultur anpasst. Auch beschreibt er [!]Jugendlich Jugendliche[!] als: „labil, unsicher, schwankend und als solche Menschen die jede Dummheit mitmachen.
Als junger Mensch kann man dieses^diesem^ Zitat nicht teilen zustimmen da die Jugend auch die Phase des Lebens sein soll um sich in einer [!]Geselschaft Gesellschaft[!] zu [!]indegrieren integrieren[!]. Somit ist es zwar richtig [!]das dass[!] man [??] als Jugendlicher unsicher ist aber diese Unsicherheit [!]benodigt benötigt[!] man um sich den richtigen Weg für Leben [!]herauszu zu suchen herauszusuchen[!]. Auch ist es [!]war wahr[!] [!]das dass[!] man als Teenager jede Dummheit [!]mit macht mitmacht[!], ob aber wenn man diese Erfahrung nie machen würde, könnte man es nie als Dummheit ansehen. Ein weiterer Aspekt ist auch die Klamottensucht, man kann es nicht als Sucht ansehen nur weil ein Jugendlicher mit der Mode geht. Es ist hingegen sehr viel schlimmer, wenn man ältere Menschen beobachtet die schon 30 Jahre [!]den seld selben denselben[!]
xsi:schemaLocation="http://www.e
<transcriber>Vera</transcriber>
<author_id>1244</author_id>
<exam_type>Koko</exam_type>
<text_language>Deutsch</text_lang
<body>
<exercise number="1"
xml:space="preserve">
„Die Jugend ist sowieso ke
ine beneidenswerte Phase“,
dieses Zitat
<correction>
<deletion>ist</deletion>
<insertion>wurde</insertion>
</correction>
in der Zeitung
<error>
<errorType eType="28 Eigennam
<originalForm>d
i
e</originalForm
<targetForm>Die</targetForm>
</error>
Zeit vom 4. Mai 2001 von
Essayist Hans Magnus Enzen
sberger geschrieben. Er be
schreibt darin
Edit
Attributes
Attribute
Value
xsi:noNamespaceSchemaL...
xsi:schemaLocation
xsi:type
INS
Desktop
DE
09:30


C:\Software\Data\xmlmind\annotations\ID1244.xml
File Select Edit Search View Tools XML Window Options Help
kolipsi_document body exercise correction insertion
ID1244.xml
Transcriber: Vera
Author ID: 1244
Exam type: Koko
Exam language: Deutsch
Exercise 1
„Die Jugend ist sowieso keine beneidenswerte Phase“, dies
Zitat ist^wurde^ in der Zeitung [!]die Die[!] Zeit vom 4. Mai 2
von Essayist Hans Magnus Enzensberger geschrieben. Er
beschreibt darin [!]das dass[!] die Jugend ein nicht zu
[!]erstrebenswerdes erstrebenswertes[!] Alter, ist, da in dies
Zeitspanne des Lebens man weder Souveränität hat, noch
der Kultur anpasst. Auch beschreibt er [!]Jugendlich
Jugendliche[!] als: „labil, unsicher, schwankend und als so
Menschen die jede Dummheit mitmachen.
Als junger Mensch kann man dieses^diesem^ Zitat nicht teilen
zustimmen da die Jugend auch die Phase des Lebens sein soll
um sich in einer [!]Geselschaft Gesellschaft[!] zu [!]indegrieren
integrieren[!]. Somit ist es zwar richtig [!]das dass[!] man [??] als
Jugendlicher unsicher ist aber diese Unsicherheit [!]benodigt
benötigt[!] man um sich den richtigen Weg für Leben
[!]herauszu zu suchen herauszusuchen[!]. Auch ist es [!]war
wahr[!] [!]das dass[!] man als Teenager jede Dummheit [!]mit
macht mitmacht[!], ob aber wenn man diese Erfahrung nie
machen würde, könnte man es nie als Dummheit ansehen. Ein
weiterer Aspekt ist auch die Klamottensucht, man kann es nicht
als Sucht ansehen nur weil ein Jugendlicher mit der Mode geht.
Es ist hingegen sehr viel schlimmer, wenn man ältere Menschen
beobachtet die schon 30 Jahre [!]den seld selben denselben[!]
ambiguous
arrow
closing
comment
correction
direct_speech
emoticon
emphasis
entity
error
footnote
footnote_marker
foreign_word
gap
greeting
hyphen
image
over
in der Zeitung
<error>
<errorType eType="28 Eigennam
<originalForm> d i e </originalForm
<targetForm> Die </targetForm>
</error>
Zeit vom 4. Mai 2001 von
Essayist Hans Magnus Enzen
sberger geschrieben. Er be
schreibt darin


C:\Software\Data\xmlmind\annotations\ID1244.xml (modified)
File Select Edit Search View Tools XML Window Options Help
kolipsi_document#body#exercise#error
ID1244.xml
[!]benodigt benötigt[!] man um sich den richtigen Weg
suchen herauszusuchen[!]. Auch ist es [!]war wahr[!]
Teenager jede Dummheit [!]mit macht mitmacht[!], ob
Erfahrung nie machen würde, könnte man es nie als
weiterer Aspekt ist auch die Klamottensucht, man kan
ansehen nur weil ein Jugendlicher mit der Mode geht.
schlimmer, wenn man ältere Menschen beobachtet di
seld selben denselben[!] „Mitst Mist“ tragen und sich
fühlen. Auch ist es notwendig, wenn man teilweise bis
Nacht[!]; eine Diskothek oder [!]ähnliches Ähnliches[!]
heutigen Kultur zu [!]witmen widmen[!] und nicht der
mehreren hundert Jahren als das [!]besste beste[!] [!]
Dagewesene[!] angesehen wird. [!]Wiederrum Wieder
merkwürdig, wenn ein [!]70ig jähriger 70-Jähriger[!] et
schreibt, da er seine Jugend in einer ganz anderen Pl
gelebt hat. Somit kann er [!]garnicht gar nicht[!] über
sprechen da seine Jugenden den Text entsprechend
Argument von H. ((Hans)) M. ((Magnus)) Enzensberg
Jugendliche sich immer das kaufen möchten [!]wass
ist in der hu heutigen Zeit auch nicht so, da ein solche
Geräten[!] und Waren auf den Markt sind [!]das dass[
besitzt[!] was ein anderer nicht hat. Somit wird die Jug
[!]Wirdschaftlichen wirtschaftlichen[!] Überschuss [!]g
[!]führ führt[!] dazu [!]das dass[!] sich auch Jugendlic
[!]übermäsig übermäßig[!] teuer waren, leisten könne
ein Jugendlicher auch [!]uberschüss Überschuss[!] vo
sich diese, durch die gute [!]Wirdschaftslage Wirtscha
[!]Weiderhin Weiterhin[!] kann eine [!]einsellne einzeln
über die [!]gesammte gesamte[!] Jugend [!]endscheid
gut oder schlecht ist, da jeder eine andere Meinung d
muss man als [!]altere ältere[!] Mensch in der Befolgerung einsehen [!]das
dass[!] sich au die [!]Geselschaft Gesellschaft[!] geändert hat und die [!] Jugen
<error>
<errorType count="2"
eType="17 insertions: other cases"/>
<errorType eType="07 sep instead of tog: other ca
<originalForm>70ig jähriger </originalForm>
<targetForm>70-Jähriger </targetForm>
</error>
etwas über die Jugend schreibt, da er s
eine Jugend in einer ganz anderen Phase
der Geschichte gelebt hat. Somit kann er
<error>
<errorType eType="08 tog instead of sep: minimal
<originalForm>garnicht </originalForm>
<targetForm>gar nicht </targetForm>
</error>
über die heutige Jugend sprechen da sei
ne Jugenden den Text entsprechend sehr l
ahm war. Ein weiter Argument von
<reduction>
<reductionType rType="02 abbreviation"/>
<reducedForm>H. </reducedForm>
<unfoldedForm>Hans </unfoldedForm>
</reduction>
</targetForm>
rtions: of double consc
</originalForm>
</targetForm>
ürdig, wenn ein
her cases"/>
instead of tog: other ca
iger </originalForm>
</targetForm>
schreibt, da er s
nz anderen Phase
at. Somit kann er
instead of sep: minimal
</originalForm>
</targetForm>
d sprechen da sei
ne Jugenden den Text entsprechend sehr l
Desktop DE 10:13


C:\Software\Data\xmlmind\annotations\ID1244.xml (modified)
File Select Edit Search View Tools XML Window Options Help
kolipsi_document#body#exercise#error
ID1244.xml
[!]benodigt benötigt[!] man um sich den richtigen Weg für Leben [!]herauszu zu suchen herauszusuchen[!]. Auch ist es [!]war wahr[!] [!]das dass[!] man als Teenager jede Dummheit [!]mit macht mitmacht[!], ob aber wenn man diese Erfahrung nie machen würde, könnte man es nie als Dummheit ansehen. Ein weiterer Aspekt ist auch die Klamottensucht, man kann es nicht als Sucht ansehen nur weil ein Jugendlicher mit der Mode geht. Es ist hingegen sehr viel schlimmer, wenn man ältere Menschen beobachtet die schon 30 Jahre [!]den seld selben denselben[!] „Mitst Mist“ tragen und sich darin auch noch wohl fühlen. Auch ist es notwendig, wenn man teilweise bis spät in die [!]Nach
Nacht[
heutig
mehre
Dagew
merkw
schreib
gelebt
sprech
Argum
Jugen
ist in d
Geräte
besitzt[!] was ein anderer nicht hat. Somit wird die Jugend von heute von einen [!]Wirdschaftlichen wirtschaftlichen[!] Überschuss [!]gebrägt geprägt[!] und dies [!]führ führt[!] dazu [!]das dass[!] sich auch Jugendliche Dinge, die 1940 noch [!]übermäsig übermäßig[!] teuer waren, leisten können. Aus diesem Grund wird ein Jugendlicher auch [!]uberschüss Überschuss[!] von Klamotten haben da er sich diese, durch die gute [!]Wirdschaftslage Wirtschaftslage[!], leisten kann. [!]Weiderhin Weiterhin[!] kann eine [!]einsellne einzelne[!] Person nicht allein über die [!]gesammte gesamte[!] Jugend [!]endscheiden entscheiden[!] ob sie gut oder schlecht ist, da jeder eine andere Meinung dazu haben kann. Auch muss man als [!]altere ältere[!] Mensch in der Befolgerung einsehen [!]das
dass[!] sich ein die [!]Geselschaft Gesellschaft[!] geändert hat und die [!] Jugen
<targetForm>je Dagewesene</targetForm>
</error>
angesehen wird.
<error>
<errorType eType="15 insertions: of double consc
<originalForm>Wiederrum</originalForm>
<targetForm>Wiederum</targetForm>
Dagewesene[!] angesehen wird. [!]Wiederrum Wiederum[!] ist es auch sehr merkwürdig, wenn ein [!]70ig jähriger 70-Jähriger[!] etwas über die Jugend schreibt, da er seine Jugend in einer ganz anderen Phase der Geschichte gelebt hat. Somit kann er [!]garnicht gar nicht[!] über die heutige Jugend sprechen da seine Jugenden den Text entsprechend sehr lahm war. Ein weiter Argument von H. ((Hans)) M. ((Magnus)) Enzensberger war [!]das dass[!] sich Jugendliche sich immer das kaufen möchten [!]wass was[!] andere haben. Die
der Geschichte gelebt hat. Somit kann er
<error>
<errorType eType="08 tog instead of sep: minimal
<originalForm>garnicht</originalForm>
<targetForm>gar nicht</targetForm>
</error>
über die heutige Jugend sprechen da sei
ne Jugenden den Text entsprechend sehr l
Edit
Attributes
INS
Desktop DE 10:13

Automatische Verarbeitung

- Lemmatisierungs- und POS-Tagging-Prozesse als Möglichkeit **der Korrektur der Transkription / man. Annotation**
- Lemmatisierung und POS-Tagging auf der Basis der **original words** und der hinzugefügten **target words**
- Alinierung: **„Zeilenausgleich“**, wenn Unterschiede bei *original words* und *target words* (v.a. Fehler bei Getrennt-/ Zusammenschreibung)
- Manuelle Lemma-/POS-Korrekturen werden ins Lexikon übernommen

Phase 1

Koko_0			KoKo_1.1		
original word	POS	Lemma	original word	POS	Lemma
,	$,	,	,	$,	,
wenn	KOUS	wenn	wenn	KOUS	wenn
ein	ART	ein	ein	ART	ein
70ig	ADJA	<unknown>	70ig___jähriger	ADJA	UNKNOWN
Jähriger	NN	<unknown>	etwas	ADV	etwas
etwas	ADV	etwas	über	APPR	über
über	APPR	über	die	ART	d
die	ART	d	Jugend	NN	Jugend
Jugend	NN	Jugend	schreibt	VVFIN	schreiben
schreibt	VVFIN	schreiben	,	$,	,
,	$,	,	da	KOUS	da

Zeilenkorrektur für orthographische Fehler- bzgl. Getrennt-/Zusammenschreibung

Phase 2

Version 1.1			Version 1.2: + targets		
original word	POS	Lemma	target word	POS	Lemma
<s>			<s>		
Wiederrum	NN	UNKNOWN	Wiederum	ADV	wiederum
ist	VAFIN	sein	ist	VAFIN	sein
es	PPER	es	es	PPER	es
auch	ADV	auch	auch	ADV	auch
sehr	ADV	sehr	sehr	ADV	sehr
merkwürdig	ADJD	merkwürdig	merkwürdig	ADJD	merkwürdig
,	$,	,	,	$,	,
wenn	KOUS	wenn	wenn	KOUS	wenn

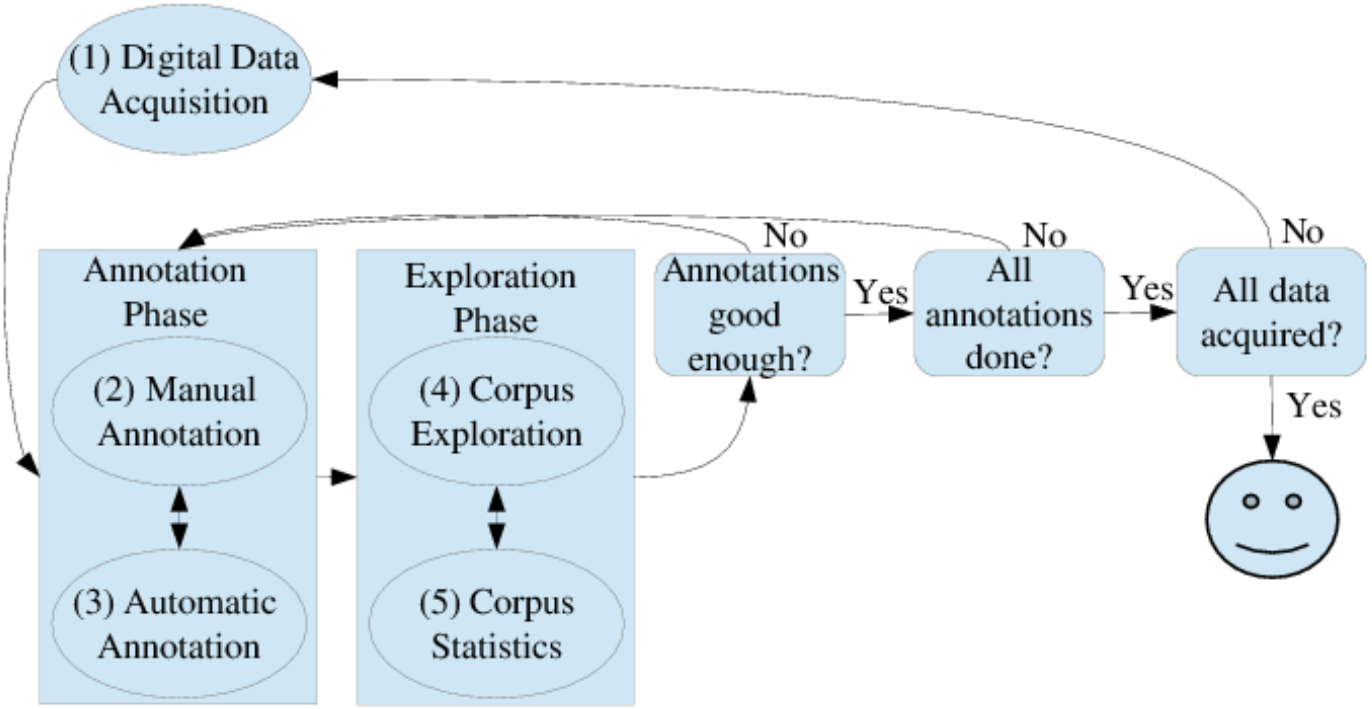
Automatische Annotation auf der Grundlage der target words

Phase 3

Version 1.2: + targets			Version 2: + manuelle Korrektur		
target word	POS	Lemma	target word	POS	Lemma
,	$,	,	,	$,	,
wenn	KOUS	wenn	wenn	KOUS	wenn
ein	ART	ein	ein	ART	ein
70-Jähriger	NN	UNKNOWN	70-Jähriger	NN	70-Jähriger
etwas	ADV	etwas	etwas	ADV	etwas
über	APPR	über	über	APPR	über
die	ART	d	die	ART	d
Jugend	NN	Jugend	Jugend	NN	Jugend
schreibt	VVFIN	schreiben	schreibt	VVFIN	schreiben
,	$,	,	,	$,	,
da	KOUS	da	da	KOUS	da

Manuelle Korrektur nicht erkannter Lemmata (OOVs) und eventueller Taggingfehler

Workflow


(1) Digital Data Acquisition
Annotation Phase
(2) Manual Annotation
(3) Automatic Annotation
Exploration Phase
(4) Corpus Exploration
(5) Corpus Statistics
Annotations good enough?
No
Yes
All annotations done?
No
Yes
All data acquired?
No
Yes

Evaluation I

Text	Tokens	Orth. Fehler	Text-länge	Satz-länge	UNKNOWNS (Prozent)			POS korrekt (Prozent)		
		%	Sätze	Worte	V1.1:	V1.2	V2	V1.1:	V1.2	V2
…										
ID1097	373	7,524	30	10,63	4,290	1,341	0,268	91,421	94,906	95,442
ID1128	626	0,926	28	19,27	2,077	0,959	0,160	97,923	98,243	98,882
ID1244	671	10,484	25	24,80	6,855	0,596	0,000	89,717	94,039	94,039
ID1491	609	2,222	30	18,00	1,806	0,657	0,000	95,895	96,552	96,716
ID2005	646	3,697	42	12,88	2,477	0,310	0,000	94,427	95,511	95,666
ID2934	632	1,667	27	20,00	1,582	0,949	0,317	95,411	96,835	96,994
…										
IDs	3557	4,4198	182	17,60	3,149	0,759	0,113	94,265	96,064	96,317
TOTAL	930241	1,78	46734	17.36	1,799	0,836	0,096	--	--	--

Evaluation II

245 Sätze (random sample) wurden evaluiert:

- POS-Informationen bzgl. unseres Goldstandards und
- Lemma-Information bzgl. der Verminderung von “unknown”-s.

Die 245 Sätze wurden weiter unterteilt in:
(1) 39 Sätze mit Änderungen im 'target level' und
(2) 206 anderen

Korpus	Size		Accuracy in %					
			KoKo 1.1		KoKo 1.2		KoKo 2	
	tok	sent	tok	sent	tok	sent	tok	sent
sample	4622	245	95.91	49.80	96.60	54.69	96.71	56.33
subdiv. 1	882	39	92.97	17.95	96.60	48.72	96.60	48.72
subdiv. 2	3740	206	96.60	55.83	96.60	55.83	96.74	57.77

Zusammenfassung

- Automatische Verarbeitung als Bestandteil der Korpusverbesserung: orthographische / Transkriptionsfehler
- Hinzufügen von target words:
  - … verbessert die POS-Tagger-Leistung teilweise erheblich, besonders bei Texten mit vielen Abweichungen von der Standardschreibung (vgl. Rehbein et al 2012).
  - … reduziert die Anzahl der UNKNOWNS automatisch.
- Manuelles Verbessern der UNKNOWNS (Hinzufügen zum Lexikon) + ggf. Verbesserung der POS-Tags
  - … verbessert in vergleichsweise geringer Zahl die POS-Tagger-Leistung
  - … reduziert weiterhin die Anzahl der UNKNOWNS

→ Sukzessive Verbesserung der Qualität der Lemma- und POS-Annotationen auch für weitere Korrektureingriffe zu erwarten.

Ausblick – Work in Progress

KoKo_3:

- Annotation des Korpus nach lexikalischen, grammatikalischen (Fehler) and textuellen Eigenschaften (Stand-off Annotation, Mmax2)
- Verwendung eines revision control system (subversion), um Fehler zu vermeiden, die bei der Annotation verschiedener Personen auftreten können
- Zugang zum Korpus über ANNIS (vgl. Zeldes 2012)