Altlitauisch Digital: Corpus des Kristijonas Donelaitis (1714–1780) [CorDon]. *Digitale Präpräsentation*

Armin Hoenen

Goethe Universität Frankfurt, Germany
Kurs: Altlitauisch

23-01-2020

Outline

Input/Output

Projektziele “vom Input zum Output”: *Sichtbarkeit, Verstehbarkeit, Weiterführung in und mit den Geisteswissenschaften*

Input - existent oder in Bearbeitung

- tiefenannotiertes Korpus
- lexikalische Ressourcen

Input/Output

Projektziele “vom Input zum Output”: *Sichtbarkeit, Verstehbarkeit, Weiterführung in und mit den Geisteswissenschaften*

Input - existent oder in Bearbeitung

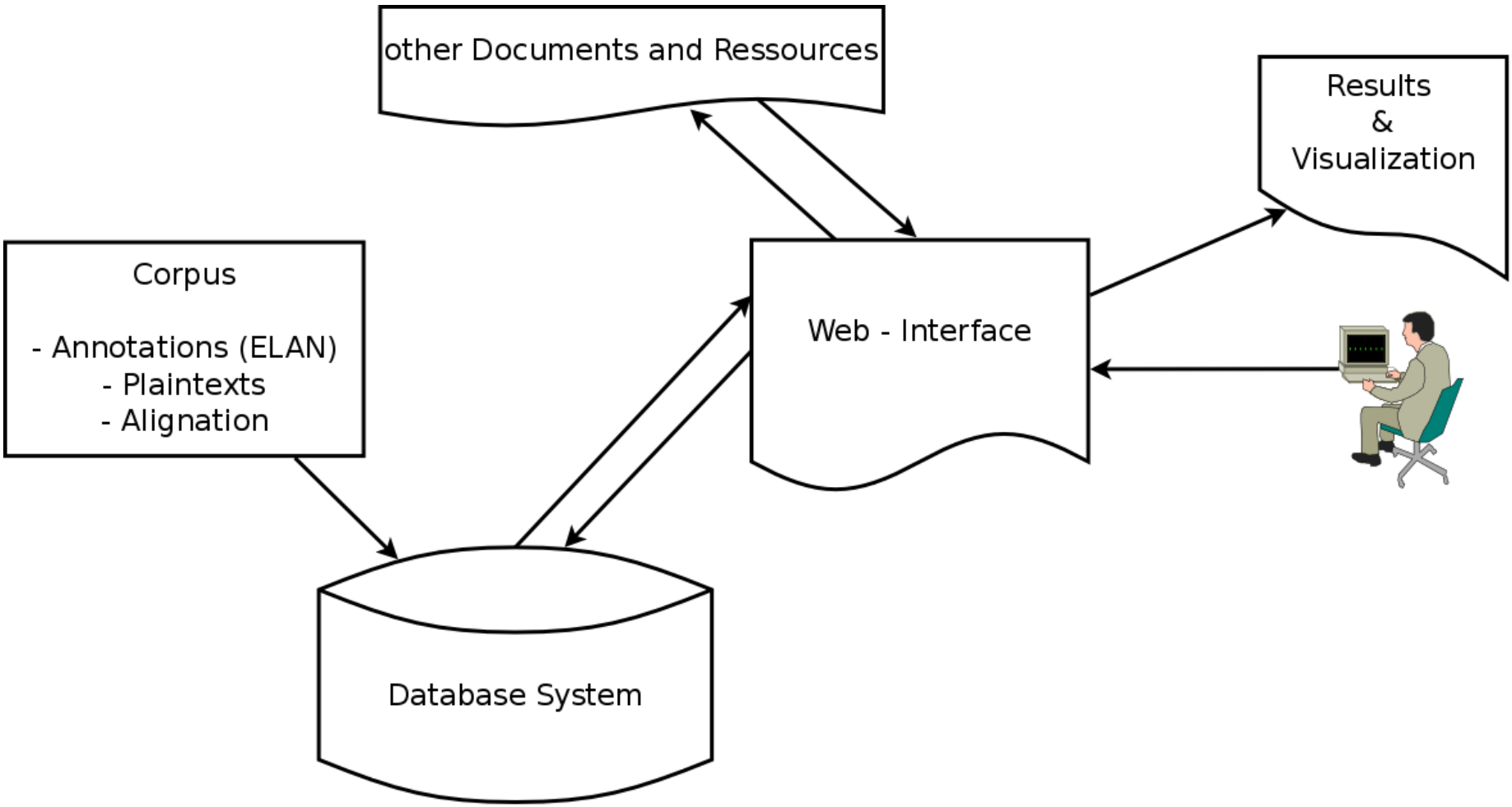
- tiefenannotiertes Korpus
- lexikalische Ressourcen

Output – to be implemented

- Web-Interface mit Suchfunktion
- Visualisierungen (Distant Reading; Moretti (2013))

Zwischen Input & Output

Technisches “Backend”: *digitale-Domäne*

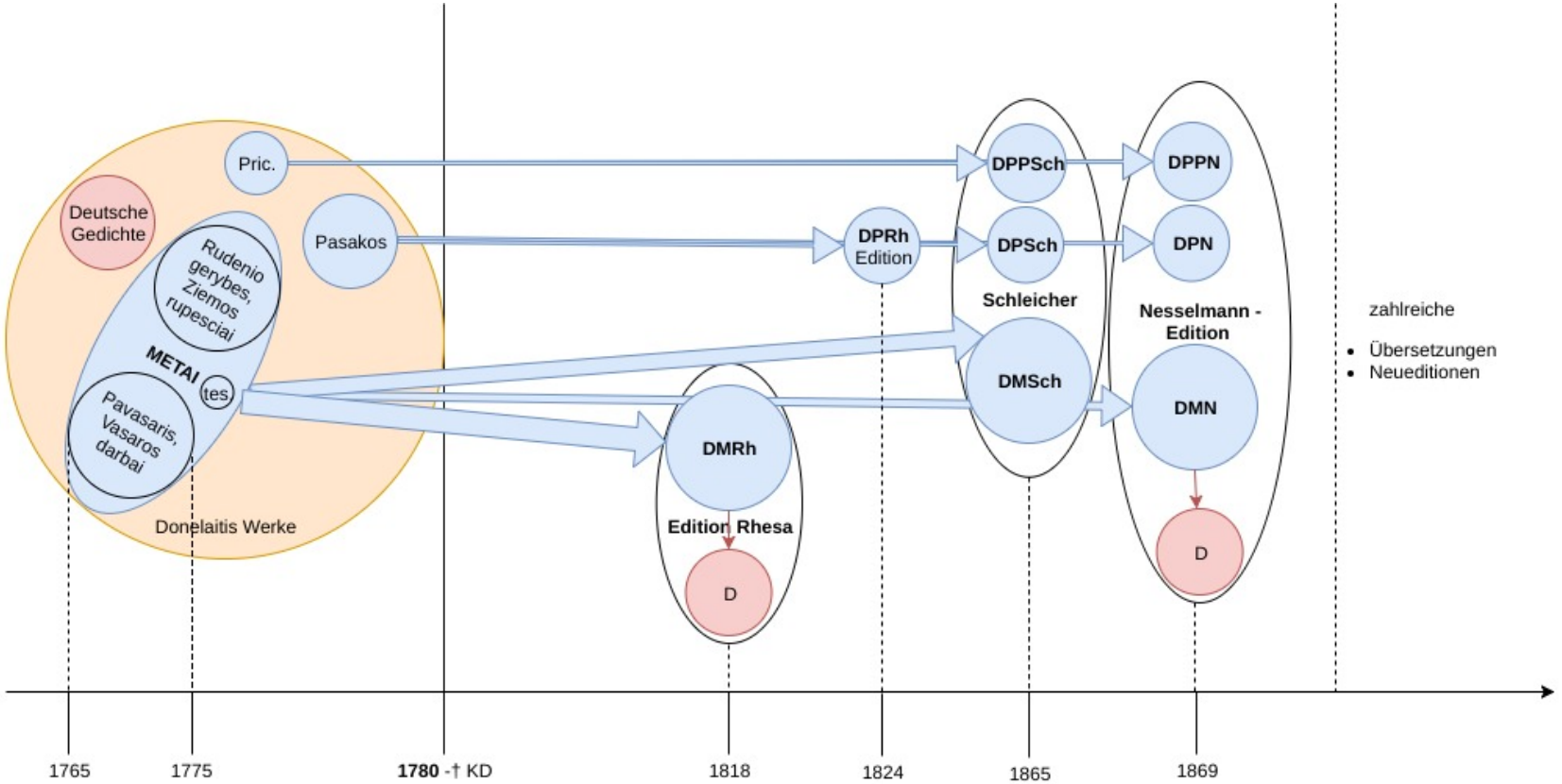


Funktionen

EINSTIEG

Graph als Einstieg

Inter- und Intratextualität



Shneiderman (1996) - “overview first [skipbar, mehrere Zugänge], zoom-in, details on demand”

Funktionen

SUCHE

Zwischen Input & Output

Suchinterface:

- Mehrere Zugänge: via
  - kanonischer Referenz

Zwischen Input & Output

Suchinterface:

- Mehrere Zugänge: via
  - kanonischer Referenz
  - grammatikalischem Filter (VERB, Imperfekt, 3. Singular)

Zwischen Input & Output

Suchinterface:

- Mehrere Zugänge: via
  - kanonischer Referenz
  - grammatikalischem Filter (VERB, Imperfekt, 3. Singular)
  - statistische Information (kostenlos): Frequenzklasse, Implikation zu Funktionswort/Inhaltswort

Zwischen Input & Output

Suchinterface:

- Mehrere Zugänge: via
  - kanonischer Referenz
  - grammatikalischem Filter (VERB, Imperfekt, 3. Singular)
  - statistische Information (kostenlos): Frequenzklasse, Implikation zu Funktionswort/Inhaltswort
  - semantische Spuren (unsupervised): Kollokationen, Nachbarn im Wortvektorraum

Zwischen Input & Output

Suchinterface:

- Mehrere Zugänge: via
  - kanonischer Referenz
  - grammatikalischem Filter (VERB, Imperfekt, 3. Singular)
  - statistische Information (kostenlos): Frequenzklasse, Implikation zu Funktionswort/Inhaltswort
  - semantische Spuren (unsupervised): Kollokationen, Nachbarn im Wortvektorraum
  - metrische Einheiten? Syntax?

Zwischen Input & Output

Suchinterface:

- Mehrere Zugänge: via
  - kanonischer Referenz
  - grammatikalischem Filter (VERB, Imperfekt, 3. Singular)
  - statistische Information (kostenlos): Frequenzklasse, Implikation zu Funktionswort/Inhaltswort
  - semantische Spuren (unsupervised): Kollokationen, Nachbarn im Wortvektorraum
  - metrische Einheiten? Syntax?
  - Alignierungsrelationen: stark, schwach, etc.

Zwischen Input & Output

Suchinterface:

- Mehrere Zugänge: via
  - kanonischer Referenz
  - grammatikalischem Filter (VERB, Imperfekt, 3. Singular)
  - statistische Information (kostenlos): Frequenzklasse, Implikation zu Funktionswort/Inhaltswort
  - semantische Spuren (unsupervised): Kollokationen, Nachbarn im Wortvektorraum
  - metrische Einheiten? Syntax?
  - Alignierungsrelationen: stark, schwach, etc.
- *simpler Default* & komplexerer, nicht-informatiker-verstehbarer *Expertenmodus*: intelligente Wildcards ($*$) und gutes Konzept für dynamische Interaktion

Zwischen Input & Output

Suchinterface:

- Mehrere Zugänge: via
  - kanonischer Referenz
  - grammatikalischem Filter (VERB, Imperfekt, 3. Singular)
  - statistische Information (kostenlos): Frequenzklasse, Implikation zu Funktionswort/Inhaltswort
  - semantische Spuren (unsupervised): Kollokationen, Nachbarn im Wortvektorraum
  - metrische Einheiten? Syntax?
  - Alignierungsrelationen: stark, schwach, etc.
- *simpler Default* & komplexerer, nicht-informatiker-verstehbarer *Expertenmodus*: intelligente Wildcards ($*$) und gutes Konzept für dynamische Interaktion
- visuelle Ästhetik sollte intuitive Bedienbarkeit fördern

Präzedenzen

P. Hinkelmanns hat eine kompakte html-Darstellung für ELAN-Daten mit dynamischen Mechanismen entwickelt (Reduktion auf weniger Ebenen).

Martynas Mažvydas MžGA ▾ ☑ Referencinis Ž originalas ☐ Transliteracija sfŽ

☐ Dabartinė nf ☑ Lema lm ☐ Lema lma ☐ Glosa (lema) glL

☐ Glosa (lema) geL ☐ Glosa (lema) D ☐ Kalba ☐ M1aL lema psL

☐ M1bF forma psF ☐ M2aL lema morf. pdL ☐ M2bF forma morf. pdF

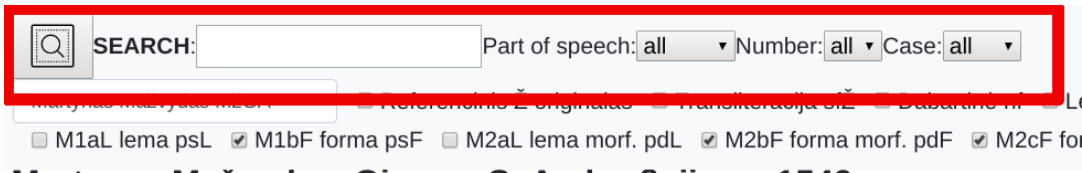
☑ M2cF forma fleks. pdlF ☐ CEN e. citatos šaltinis

MžGA 1,1	Gieſme	Schwenta	Am=|| braſʒeijaus
Referencinis Ž originalas	Gieſme	S.	Am=||braſʒeijaus
Lema lm	giesmė	šventas	Ambraziejus
M2cF forma fleks. pdlF	Sg_Nom_(e)	[Pos_Masc_Sg_Gen]	Sg_Gen

MžGA 1,2	,	bey	Schwenta	Au-||guſtina
Referencinis Ž originalas	,	bey	S.	Au-||guſtina
Lema lm		bei	šventas	Augustinas
M2cF forma fleks. pdlF		(ey)	[Pos_Masc_Sg_Gen]	Sg_Gen_(a)

Expansion I – Suche

Suche



Martynas Mažvydas, Giesme S. Ambraßeijaus, 1549

MžGA 1,1	Gieſme	Schwenta	Am=|| braſʒeijaus
Referencinis Ž originalas	Gieſme	S.	Am=||braſʒeijaus
Lema lm	giesmė	šventas	Ambraziejus
M1bF forma psF	NA	[ADJ]	NT
M2bF forma morf. pdF	ė_Fem	[a]	ju_Masc
M2cF forma fleks. pdlF	Sg_Nom_(e)	[Pos_Masc_Sg_Gen]	Sg_Gen

MžGA 1,2	,	bey	Schwenta	Au-||guſtina
Referencinis Ž originalas	,	bey	S.	Au-||guſtina
Lema lm		bei	šventas	Augustinas
M1bF forma psF	$,	KO	[ADJ]	NT
M2bF forma morf. pdF		-	[a]	a_Masc
M2cF forma fleks. pdlF		(ey)	[Pos_Masc_Sg_Gen]	Sg_Gen_(a)

MžGA 1,3	,	kurę	wadin:	Te
Referencinis Ž originalas	,	kurę	wadin:	Te
Lema lm		kuris	vadinti	tū
M1bF forma psF	$,	PKREL	V	PPER
M2bF forma morf. pdF		jo	a-Pres	
M2cF forma fleks. pdlF		Fem_Sg_Acc_(ę)	Ind_Pres_3_(-)	Sg_Acc

Expansion I – Suche

Search-Result

SEARCH: Schwenta,Schwentas Part of speech: all Number: all Case: all

Martynas Mažvydas MžGA ☑ Referencinis Ž originalas ☐ Transliteracija sfŽ ☐ Dabartinė nf ☑ Lema lm ☐ Lema lma ☐ Glosa (lema) glL ☐ Glosa (lema) geL ☐ Glosa

☐ M1aL lema psL ☑ M1bF forma psF ☐ M2aL lema morf. pdL ☑ M2bF forma morf. pdF ☑ M2cF forma fleks. pdlF ☐ CEN e. citatos šaltinis

Martynas Mažvydas, Giesme S. Ambraßeijaus, 1549

MžGA 1,1	Gieſme	Schwenta	Am=|| braſʒeijaus
Referencinis Ž originalas	Gieſme	S.	Am=||braſʒeijaus
Lema lm	giesmė	šventas	Ambraziejus
M1bF forma psF	NA	[ADJ]	NT
M2bF forma morf. pdF	ė_Fem	[a]	ju_Masc
M2cF forma fleks. pdlF	Sg_Nom_(e)	[Pos_Masc_Sg_Gen]	Sg_Gen

MžGA 1,2	,	bey	Schwenta	Au-||guſtina
Referencinis Ž originalas	,	bey	S.	Au-||guſtina
Lema lm		bei	šventas	Augustinas
M1bF forma psF	$,	KO	[ADJ]	NT
M2bF forma morf. pdF		-	[a]	a_Masc
M2cF forma fleks. pdlF		(ey)	[Pos_Masc_Sg_Gen]	Sg_Gen_(a)

MžGA 4,37	Schwentas	,	Schwentas	,
Referencinis Ž originalas	Schwe<n>tas	/	Schwentas	/
Lema lm	šventas		šventas	
M1bF forma psF	ADJP	$,	ADJP	$,
M2bF forma morf. pdF	a		a	
M2cF forma fleks. pdlF	Pos_Masc_Sg_Nom		Pos_Masc_Sg_Nom	

Alignment

ALIGNMENT

Verszeilen Text-Alignment – Overview (Expansion II)

Shneiderman's (1996) visual information seeking mantra

Overview first, zoom-in, then details on-demand

Overview:

Kristjonas Donelaitis, Nesselmann 1869

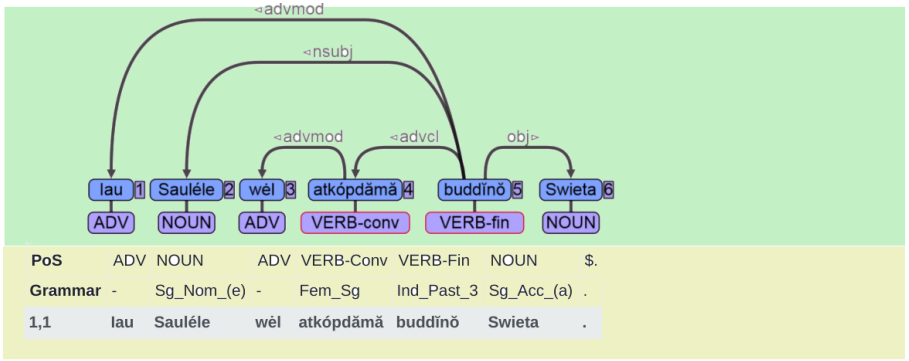
1,1	lau	Sauléle	wėl	atkópdămă	buddĩnŏ	Swieta	.
Grammar	-	Sg_Nom_(e)	-	Fem_Sg	Ind_Past_3	Sg_Acc_(a)	.
PoS	ADV	NOUN	ADV	VERB-Conv	VERB-Fin	NOUN	$.

Kristjonas Donelaitis, Nesselmann 1869

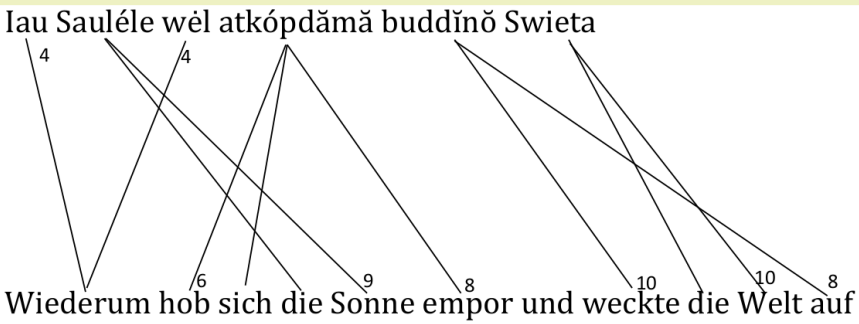
1,1	Wiederum	hob	sich	die	Sonne	empor	und	weckte	die	Welt	auf	.
Grammar	-	Ind_Past_3	refl	Nom_Sg	Nom_Sg	-	-	Ind_Past_3	Acc_Sg	Acc_Sg	-	.
PoS	ADV	VVFIN	PRF	ART	NN	ADV	KON	VVFIN	ART	NN	ADV	$.

Verszeilen Annotationsalignment (Expansion III)

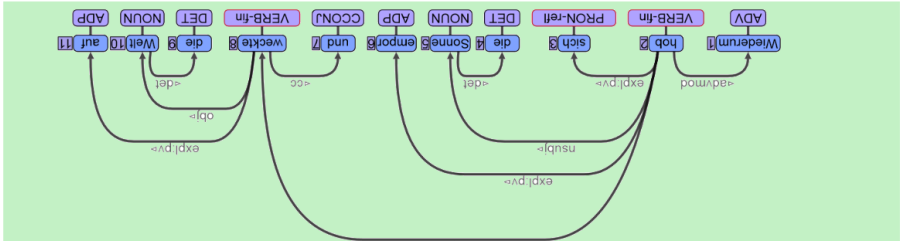
Zoom-in (durch Zeilenwahl)



PoS	ADV	NOUN	ADV	VERB-Conv	VERB-Fin	NOUN	$.
Grammar	-	Sg_Nom_(e)	-	Fem_Sg	Ind_Past_3	Sg_Acc_(a)	.
1,1	Iau	Sauléle	wél	atkópdămă	buddĭnŏ	Swieta	.

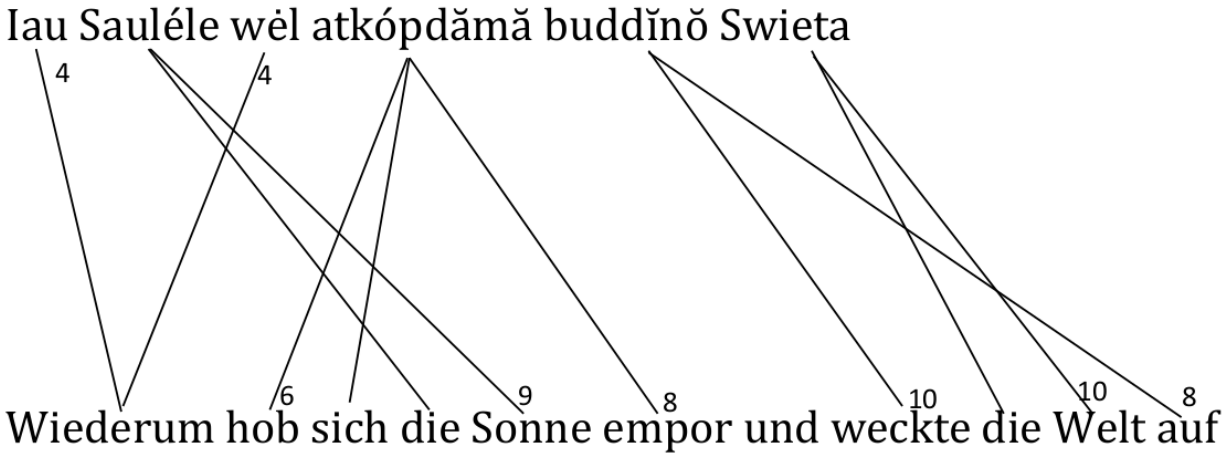


1,1	Wiederum	hob	sich	die	Sonne	empor	und	weckte	die	Welt	auf	.
Grammar	-	Ind_Past_3	refl	Nom_Sg	Nom_Sg	-	-	Ind_Past_3	Acc_Sg	Acc_Sg	-	.
PoS	ADV	VVFIN	PRF	ART	NN	ADV	KON	VVFIN	ART	NN	ADV	$.



Verszeilen Annotationsalignment (Expansion III)

details on-demand: dynamische Interaktion mit Ebenentausch,
Token Level Alignment



Differenz im morphologischen Profil

Morphologisches Profil, nach Sylak-Glassmann et al. (2015)

Dimension	*Features*	*Semantic Basis*
Aktionsart	ACCMP, ACH, ACTY, ATEL, DUR, DYN, PCT, SEMEL, STAT, TEL	Cable (2008), Vendler (1957), Comrie (1976a)
Animacy	ANIM, HUM, INAN, NHUM	Yamamoto (1999), Comrie (1989)
Aspect	HAB, IPFV, ITER, PFV, PRF, PROG, PROSP	Klein (1994)
Case	ABL, ABS, ACC, ALL, ANTE, APPRX, APUD, AT, AVR, BEN, CIRC, COM, COMPV, DAT, EQU, ERG, ESS, FRML, GEN, INS, IN, INTER, NOM, NOMS, ON, ONHR, ONVR, POST, PRIV, PROL, PROPR, PROX, PRP, PRT, REM, SUB, TERM, VERS, VOC	Blake (2001), Radkevich (2010)
Comparison	AB, CMPR, EQT, RL, SPRL	Cuzzolin and Lehmann (2004)
Definiteness	DEF, INDEF, NSPEC, SPEC	Lyons (1999)
Deixis	ABV, BEL, DIST, EVEN, MED, NVIS, PROX, REF1, REF2, REM, VIS	Bhat (2004), Bliss and Ritter (2001)
Evidentiality	ASSUM, AUD, DRCT, FH, HRSY, INFER, NFH , NVSEN, QUOT, RPRT, SEN	Aikhenvald (2004)
Finiteness	FIN, NFIN	Binary finite vs. nonfinite
Gender+	BANTU1-23, FEM, MASC, NAKH1-8, NEUT	Corbett (1991)
Info. Structure	FOC, TOP	Lambrecht (1994)
Interrogativity	DECL, INT	Binary declarative vs. interrogative
Mood	ADM, AUNPRP, AUPRP, COND, DEB, IMP, IND, INTEN, IRR, LKLY, OBLIG, OPT, PERM, POT, PURP, REAL, SBJV, SIM	Palmer (2001)
Number	DU, GPAUC, GRPL, INVN, PAUC, PL, SG, TRI	Corbett (2000)
Parts of Speech	ADJ, ADP, ADV, ART, AUX, CLF, COMP, CONJ, DET, INTJ, N, NUM, PART, PRO, V, V.CVB, V.MSDR, V.PTCP	Croft (2000), Haspelmath (1995)
Person	0, 1, 2, 3, 4, EXCL, INCL, OBV, PRX	Conventional person, obviation and clusivity
Polarity	NEG, POS	Binary positive vs. negative
Politeness	AVOID, COL, FOREG, FORM, FORM.ELEV, FORM.HUMB, HIGH, HIGH.ELEV, HIGH.SUPR, INFM, LIT, LOW, POL	Brown and Levinson (1987), Comrie (1976b)
Possession	ALN, NALN, PSSD, PSSPNO+	Type of possession, characteristics of possessor
Switch-Reference	CN-R-MN+, DS, DSADV, LOG, OR, SEQMA, SIMMA, SS, SSADV	Stirling (1993)
Tense	1DAY, FUT, HOD, IMMED, PRS, PST, RCT, RMT	Klein (1994), ?
Valency	DITR, IMPRS, INTR, TR	Number of verbal arguments from zero to three
Voice	ACFOC, ACT, AGFOC, ANTIP, APPL, BFOC, CAUS, CFOC, DIR, IFOC, INV, LFOC, MID, PASS, PFOC, RECP, REFL	Klaiman (1991)

Differenz im morphologischen Profil

Morphologisches Profil

NUMERUS

- DE - SG, PL (@N,@V,@ADJ)
- OLT - SG, DU, PL (@N,@V,@ADJ)
- LT - SG, PL (@N,@V,@ADJ)

CASE

...

Strategien, Wahrscheinlichkeiten:

FEAT	**Type**	**Bsp.**	**Prob.**
OLT NUM **DU**	lex. Erweiterung	*zu zweit*	0.6
	Informationsverlust	-	0.3
	lex. Ersetzung	*das Paar*	0.05
	anderes	z.B. Archaismus	0.05

¿Signifikante Unterschiede zwischen DE und LT in der Übersetzungskompensation (z.B. in Bezug auf Archaismen)?

Ressourcen – (Alt)Litauisch

ein Ausschnitt - moderne Sprache (Rückanpassung?)

- M OPUS: Paralleltext: LIT - DE ca. 80.000 Token

u.a. WALS

Ressourcen – (Alt)Litauisch

ein Ausschnitt - moderne Sprache (Rückanpassung?)

- M OPUS: Paralleltext: LIT - DE ca. 80.000 Token
- M Universal Dependencies: 75K Token syntaktisch annotiert (Treebank)

u.a. WALS

Ressourcen – (Alt)Litauisch

ein Ausschnitt - moderne Sprache (Rückanpassung?)

- M OPUS: Paralleltext: LIT - DE ca. 80.000 Token
- M Universal Dependencies: 75K Token syntaktisch annotiert (Treebank)
- A Titus SLIEKKAS

u.a. WALS

Ressourcen – (Alt)Litauisch

ein Ausschnitt - moderne Sprache (Rückanpassung?)

- M OPUS: Paralleltext: LIT - DE ca. 80.000 Token
- M Universal Dependencies: 75K Token syntaktisch annotiert (Treebank)
- A Titus SLIEKKAS
- A litauische Institute

u.a. WALS

Ressourcen – (Alt)Litauisch

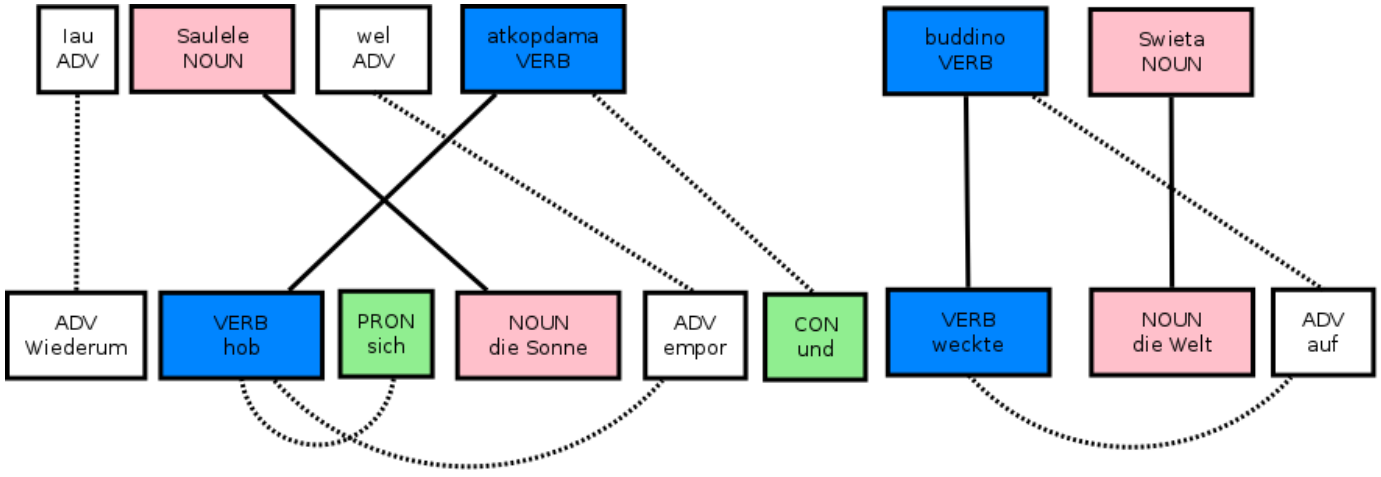
ein Ausschnitt - moderne Sprache (Rückanpassung?)

- M OPUS: Paralleltext: LIT - DE ca. 80.000 Token
- M Universal Dependencies: 75K Token syntaktisch annotiert (Treebank)
- A Titus SLIEKKAS
- A litauische Institute
- A Bibel (1735), Christodouloupoulos & Steedman (2014)

u.a. WALS

Verszeilen Annotationsalignment (Expansion III)

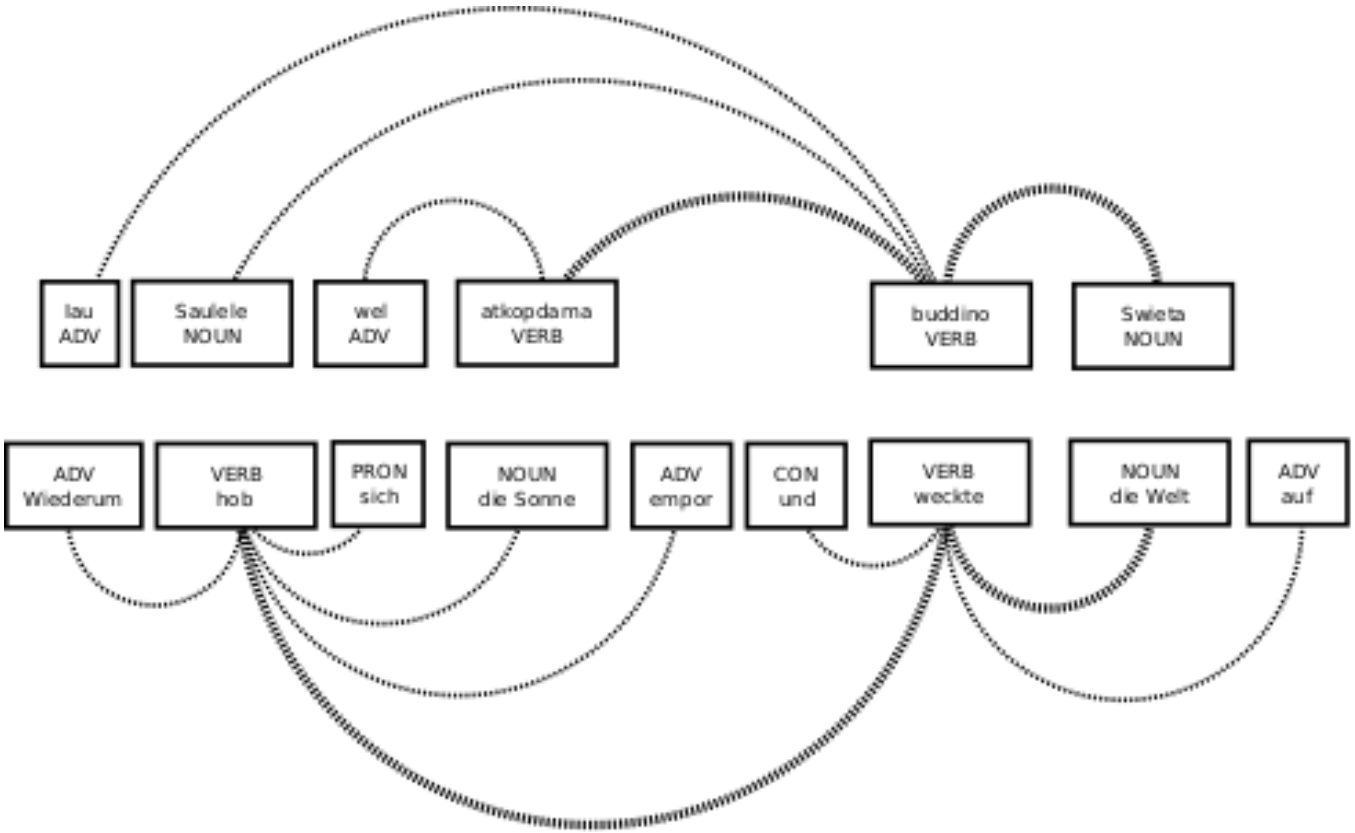
details on-demand: PoS Level Alignment



durch Mapping wie denen des Universal Tagset (Petrov et al. 2012)

Verszeilen Annotationsalignment (Expansion III)

details on-demand: Syntaktische Ebene, Alignment



mit Universal Dependencies und statistischem Vergleich

Alignment – Technisch

Human-in-the-loop kontrolliertes semi-automatisches Alignment.

- via Maschineller Übersetzung (MT)

Alignment – Technisch

Human-in-the-loop kontrolliertes semi-automatisches Alignment.

- via Maschineller Übersetzung (MT)
- GIZA++ und fastAlign, Sprachmodelle
  - externe Trainingsdaten, parallele Bibel, OPUS, SLIEKKAS

Alignment – Technisch

Human-in-the-loop kontrolliertes semi-automatisches Alignment.

- via Maschineller Übersetzung (MT)
- GIZA++ und fastAlign, Sprachmodelle
  - externe Trainingsdaten, parallele Bibel, OPUS, SLIEKKAS
- zusätzlich lexikalische Ressourcen: Alignmenstärke

Alignment – Technisch

Human-in-the-loop kontrolliertes semi-automatisches Alignment.

- via Maschineller Übersetzung (MT)
- GIZA++ und fastAlign, Sprachmodelle
  - externe Trainingsdaten, parallele Bibel, OPUS, SLIEKKAS
- zusätzlich lexikalische Ressourcen: Alignmenstärke
- POSTILLEN - Modellierung (Typologie)

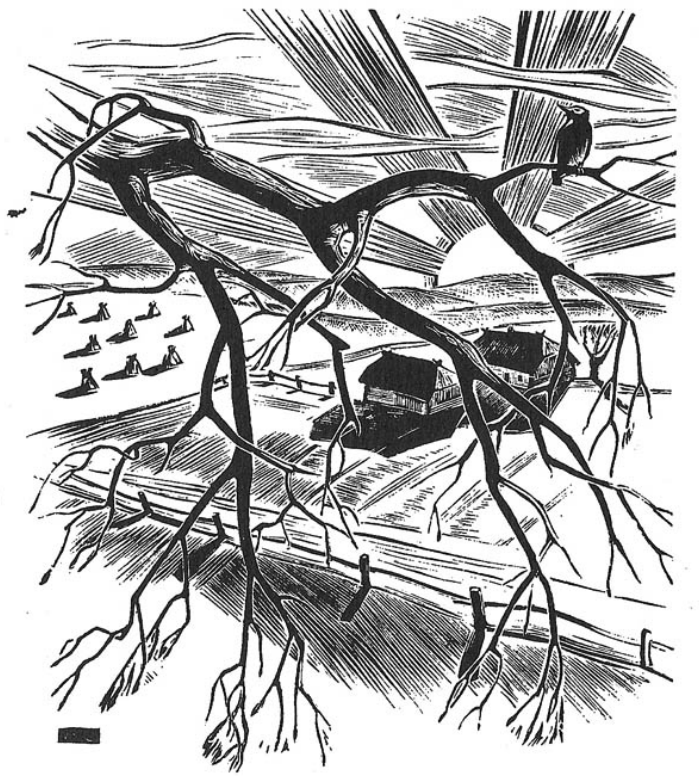
Alignment – Technisch

Human-in-the-loop kontrolliertes semi-automatisches Alignment.

- via Maschineller Übersetzung (MT)
- GIZA++ und fastAlign, Sprachmodelle
  - externe Trainingsdaten, parallele Bibel, OPUS, SLIEKKAS
- zusätzlich lexikalische Ressourcen: Alignmenstärke
- POSTILLEN - Modellierung (Typologie)
- POSTILLEN - Evaluation und technische Entwicklung - Automatisierung

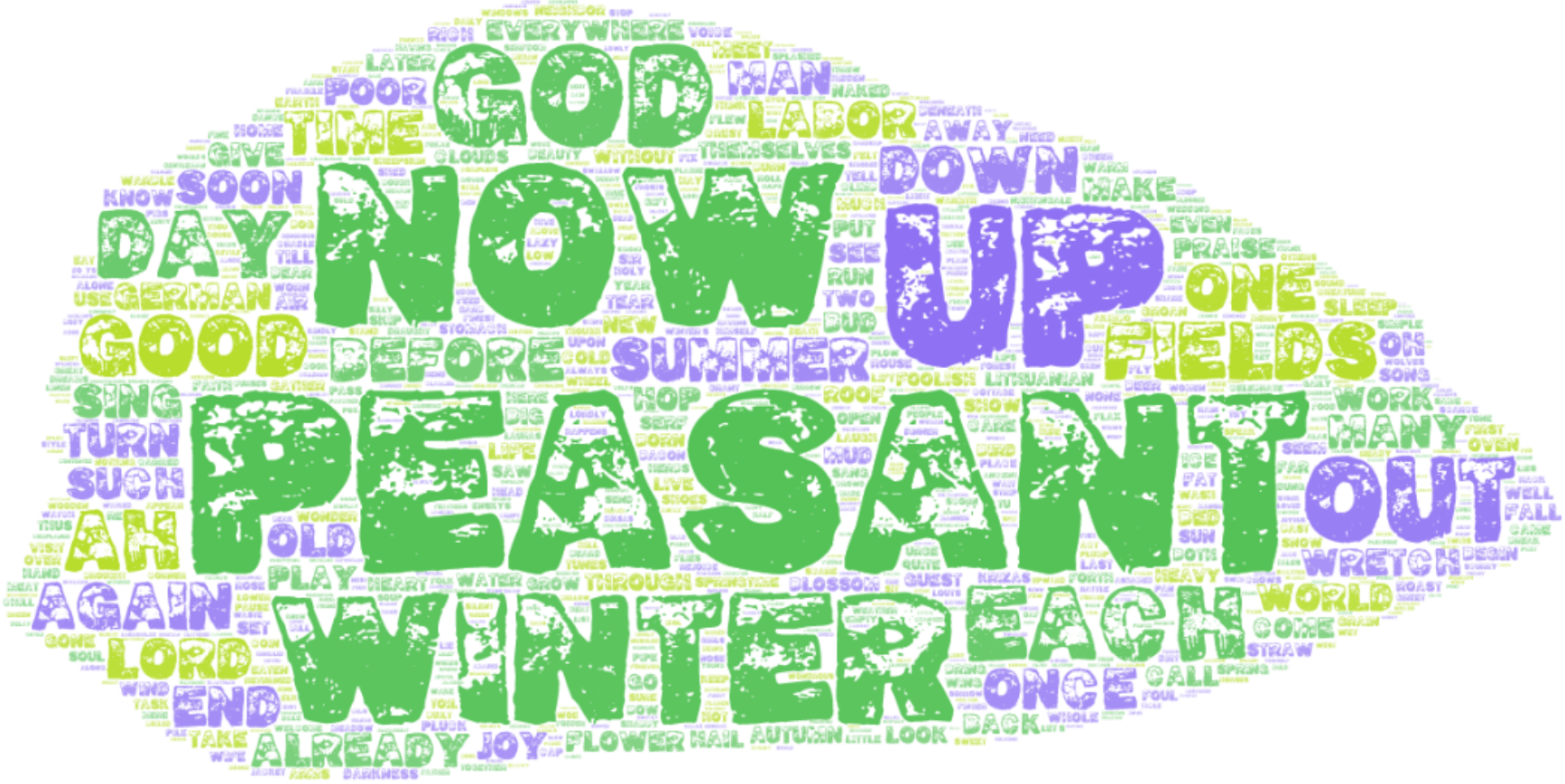
Visualisierung

VISUALISIERUNG



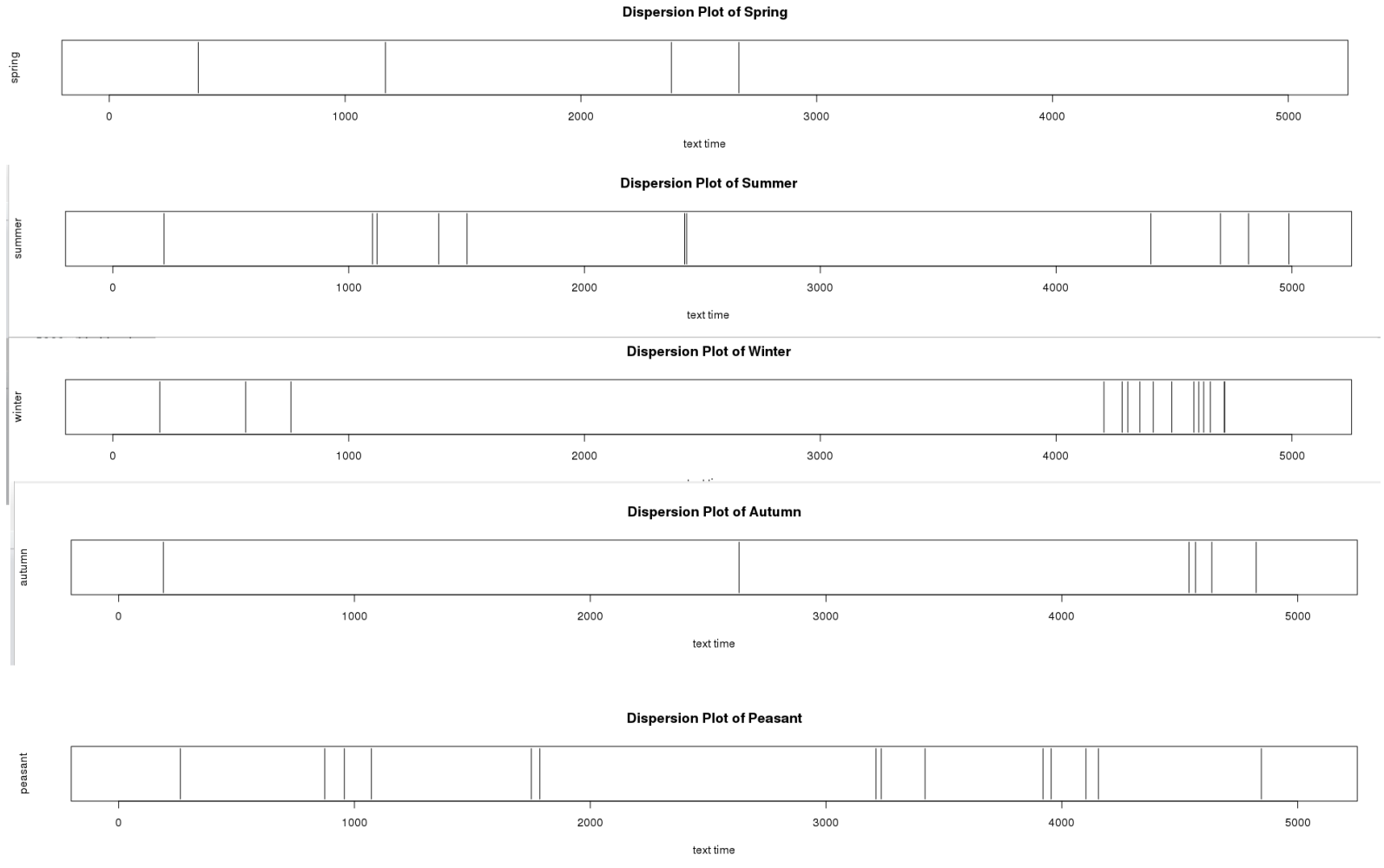
Visualisierung – schnell & existent

Aus den vorh. Ressourcen & Alignments, Statistische Graphen: Zipf Curve, Vokabelreichtum etc.



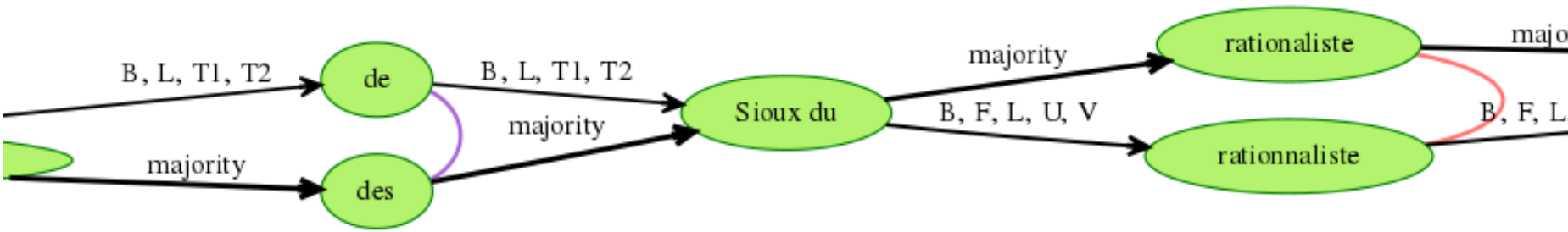
Visualisierung – schnell & existent

Dispersion Plot



Visualisierung – schnell & existent

Variantengraph, zwischen Versionen, sowie cross-linguistisch



Es wurde präsentiert:

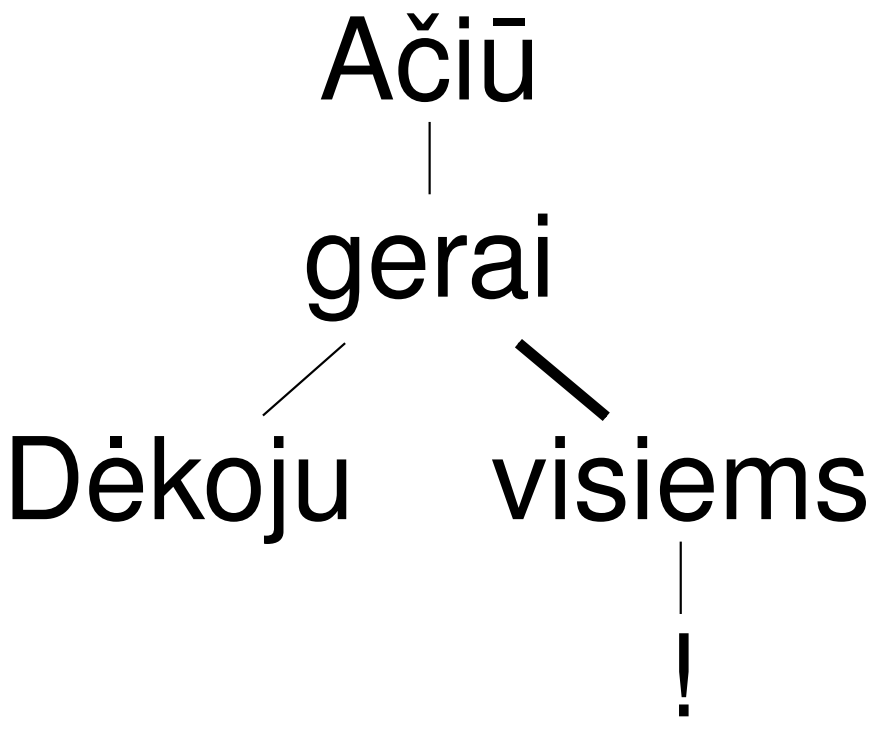
Technische Pläne

Präsentation, Suche, Alignment, Visualisierungen

& Konzepte

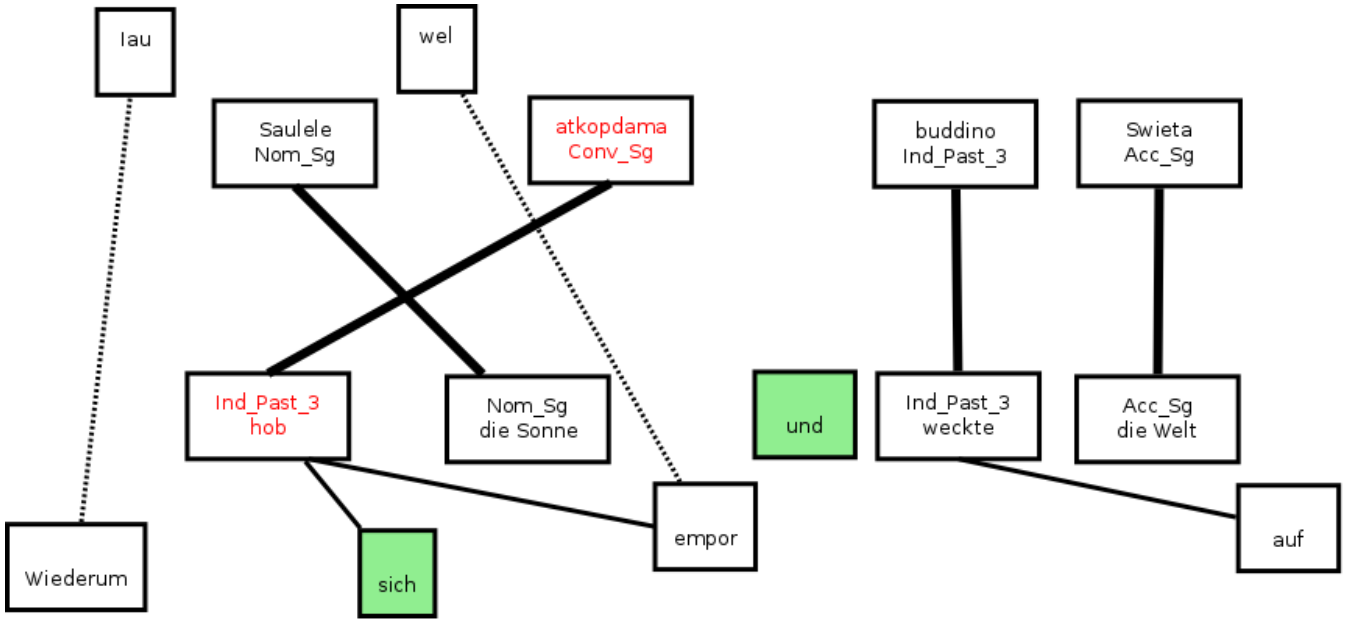
technisches Backend, Formalisierung, Evaluation

eBye



Verszeilen Annotationsalignment (Expansion III)

details on-demand: Morphologisches Alignment



viel theoretische Arbeit nötig