

**Möglichkeiten der computergestützten
Erstellung und Analyse von Korpora
gesprochener Sprache**

Gliederung

-Projektrahmen und Projekt

- EXMARaLDA Architektur und Werkzeuge
- Drei SFB-Korpora
- Demonstration „Skandinavische Semikommunikation“
- Aufarbeitung der Korpora
- Korpus-Fertigstellung
- Neue Möglichkeiten -> neue Probleme
- Anleitung zum Arbeit-sparen
- Plädoyer für bessere Datenarbeit

Projektrahmen: SFB 538 „Mehrsprachigkeit“

3+1 Projektbereiche:

Spracherwerb (E) Individueller Spracherwerb und Aspekte der Sprachverarbeitung.

Mehrsprachige Kommunikation (K) Wesen, Möglichkeiten und Risiken mehrsprachiger Kommunikation

Historische Mehrsprachigkeit (H) Rolle des Sprachwandels unter den Bedingungen der Mehrsprachigkeit und des Sprachkontakts

Neu: Transferprojekte (T) Umsetzung der Ergebnisse in praxisrelevante Anwendungen

Projektrahmen: SFB 538 „Mehrsprachigkeit“

Deutsch
(Alt-) Französisch
Portugiesisch
(Alt-) Schwedisch
Altnordisch/Isländisch
(Alt-) Dänisch
(Hiberno-) Englisch
Katalanisch
Japanisch
Türkisch
DGS
Färöisch
Brasilianisch
Baskisch
Spanisch

Projektrahmen: Projekte Z2 & C2

Z2: „Computergestützte Erfassungs- und Analysemethoden multilingualer Daten“

Schafft methodische und technologische Grundlagen für den Computereinsatz in der Mehrsprachigkeitsforschung

Projektrahmen: Projekte Z2 & C2

Z2: „Computergestützte Erfassungs- und Analysemethoden multilingualer Daten“

Schafft methodische und technologische Grundlagen für den Computereinsatz in der Mehrsprachigkeitsforschung

C2: „Nachhaltigkeit linguistischer Daten“

Nachhaltige Verfügbarkeit der Daten der beteiligten SFBs, Lösungen mit exemplarischem Charakter.

Verbundprojekt zusammen mit den SFBs

441 „Linguistische Datenstrukturen“, Tübingen und

632 „Informationsstruktur“, Potsdam

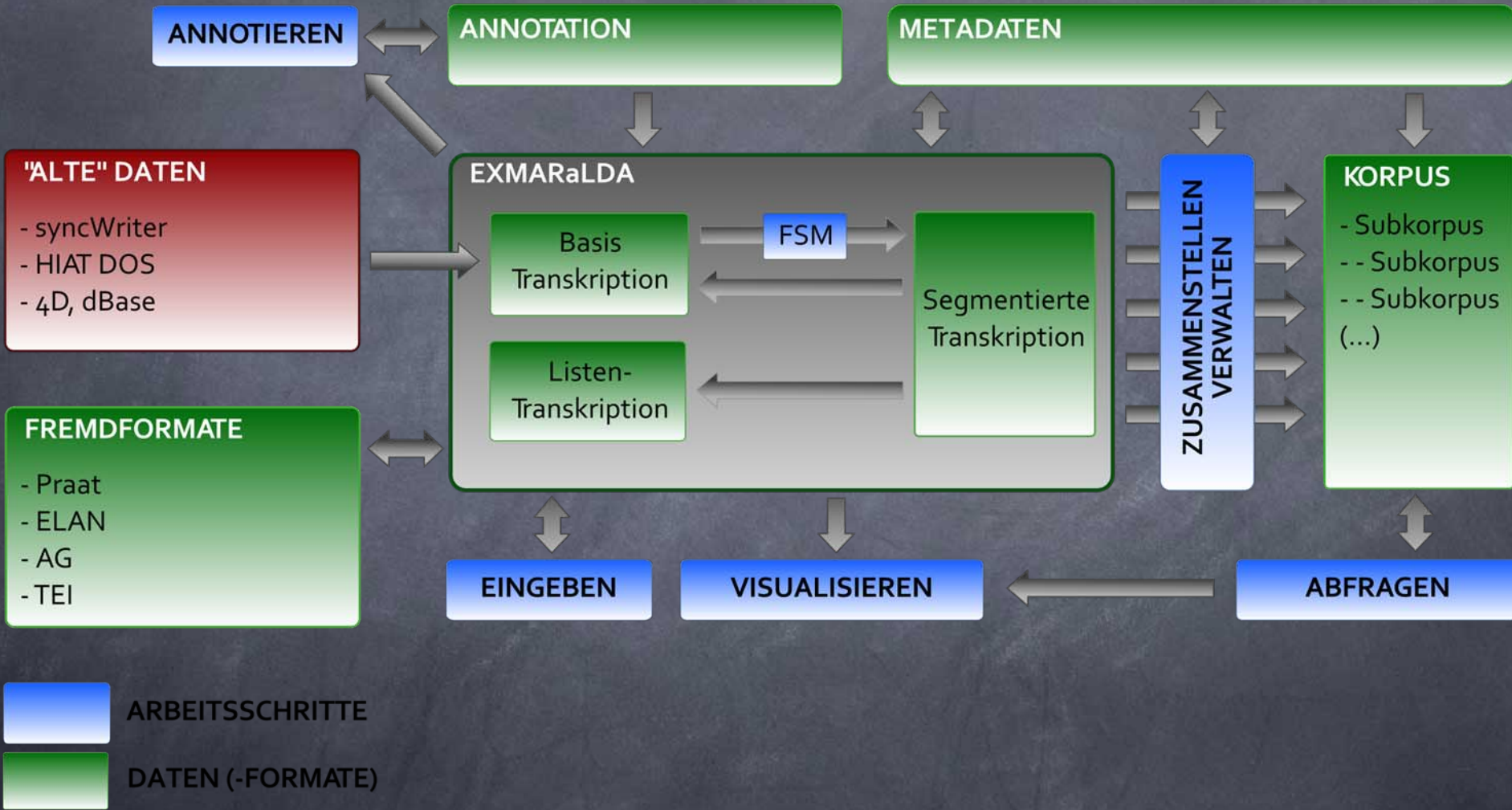
Gliederung

- Projektrahmen und Projekt
- EXMARaLDA Architektur und Werkzeuge**
- Drei SFB-Korpora
- Demonstration „Skandinavische Semikommunikation“
- Aufarbeitung der Korpora
- Korpus-Fertigstellung
- Neue Möglichkeiten -> neue Probleme
- Anleitung zum Arbeit-sparen
- Plädoyer für bessere Datenarbeit

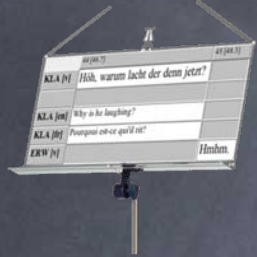
EXMARaLDA: Architektur und Werkzeuge

Extensible Markup Language for Discourse Annotation

EXMARaLDA-Systemarchitektur



Werkzeuge



Partitur-Editor



COMA (Corpus-Manager)



EXAKT (EXMARaLDA Analyse und Konkordanz-Tool)

Werkzeuge: Partitur-Editor

Transkription/Import, Bearbeitung, Segmentierung, Export, Visualisierung, Audio/Video-Verknüpfung

The screenshot displays the EXMARaLDA Partitur-Editor 1.3.4 interface. The main window shows a transcription table with columns for time (19, 20, 21, 22, 23, 24, 25) and rows for different linguistic levels (NN, [nsp], Sm205, Sm206, Dm206). The Dm206 [en] row contains the English transcription: "And it will be Poul Nyrup Rasmussen who will form a minority government, but he must get support from the Socialist People's Party." The interface also includes an Audio/Video panel showing a video of a chef and a Keyboard panel with a specialized layout for Germanic languages.

	19	20	21	22	23	24	25
NN [v]							
[nsp]							
Sm205 [v]							
Sm205 [en]							
Sm206 [v]							
Sm206 [k]							
Sm206 [en]							
Dm206 [v]	ering.	Og det vil	blive	Poul Nyrup Rasmussen som vil	bilda	en mindretalsregering, men han skal have	• stød • fra Socialistisk Folk
Dm206 [cs]					s		s
Dm206 [k]			blive				stöt
Dm206 [en]	ent.	And it will be Poul Nyrup Rasmussen who will form a minority government, but he must get support from the Socialist People's Party.					



Werkzeuge: Coma (Corpus Manager)

Sammeln von Transkriptionen zu Korpora, Verwaltung von Metadaten, Zusammenstellen von Untersuchungskorpora

Coma2 | CoMa_SKOBI_Datenbank.xml | Corpus: E5 Korpus

Datei Werkzeuge Hilfe

Korpus Daten Korpus-Korb Einstellungen

Filter
SKOBI Invert Add Active
F - Fertig Invert Add Active
Alle Filter löschen 79 Kommunikationen

Kommunikation

S	Name	Var
<input checked="" type="checkbox"/>	0736	EFE04 tk - Maulwurf
<input type="checkbox"/>	0826	EFE04 tk - Maulwurf
<input type="checkbox"/>	0693	EFE05 dt - Bildbeschreibung
<input type="checkbox"/>	0694	EFE05 dt - Bildbeschreibung
<input type="checkbox"/>	0697	EFE05 dt - Bildbeschreibung
<input type="checkbox"/>	0689	EFE05 dt - Bildbeschreibung
<input type="checkbox"/>	0698	EFE05 dt - Bildbeschreibung
<input type="checkbox"/>	0649	EFE05 tk - Bildbeschreibung
<input type="checkbox"/>	0803	EFE05 tk - Bildbeschreibung
<input type="checkbox"/>	0740	EFE05 tk - Bildbeschreibung
<input type="checkbox"/>	0734	EFE05 tk - Bildbeschreibung
<input type="checkbox"/>	0898	EFE05 tk - Bildbeschreibung
<input type="checkbox"/>	0661	EFE07 dt - Kennenlerngespräch
<input type="checkbox"/>	0687	EFE07 dt - Kennenlerngespräch
<input type="checkbox"/>	0613	EFE07 dt - Kennenlerngespräch
<input type="checkbox"/>	0704	EFE07 dt - Kennenlerngespräch
<input type="checkbox"/>	0760	EFE07 dt - Kennenlerngespräch
<input type="checkbox"/>	0819	EFE07 dt - Kennenlerngespräch
<input type="checkbox"/>	0757	EFE07 dt - Kennenlerngespräch
<input type="checkbox"/>	0750	EFE07 dt - Kennenlerngespräch
<input type="checkbox"/>	0632a/b	EFE07 tk - Kennenlerngespräch
<input type="checkbox"/>	0611	EFE07 tk - Kennenlerngespräch
<input type="checkbox"/>	0605a	EFE07 tk - Kennenlerngespräch
<input type="checkbox"/>	0654	EFE07 tk - Kennenlerngespräch
<input type="checkbox"/>	0658	EFE07 tk - Kennenlerngespräch

Aktion

Eigenschaften Sprecher anzeigen XML anzeigen *cmd.saveXML

Communication 0736

Id IDCommunication1091

Description (Communication)

Thema Inhalt

Projekt

Konvertierung

KassettenNr

FileMakerID

Familie (Pseudo)

Familie (Pseudo)

Diskursmodus

DOrt räuml.

Bearbeitungsstand

Location

City

Country

PStart

Description (Location)

Language(s)

Language

Language Code

Description (Language)

Language

Personen Transkriptionen Aufnahmen

Angezeigter Key aus Description

...	Sigle	Var
<input type="checkbox"/>	Br	Birgül
<input type="checkbox"/>	Eme	Emel
<input type="checkbox"/>	Mut12	Mütter
<input type="checkbox"/>	Ken	Kenan
<input type="checkbox"/>	Ann	Anna
<input type="checkbox"/>	Üns	Ünsal
<input checked="" type="checkbox"/>	Fer1	Ferda
<input type="checkbox"/>	Cev	Cevat
<input checked="" type="checkbox"/>	Nes	Nesli
<input type="checkbox"/>	Mut1	Mütter
<input type="checkbox"/>	Hai	Halime
<input type="checkbox"/>	Sam	Samet
<input type="checkbox"/>	Oyc	Oycan
<input type="checkbox"/>	Kam	Kamil
<input type="checkbox"/>	Mus	Musa
<input type="checkbox"/>	Vat1	Vater
<input type="checkbox"/>	Mut211	Mütter
<input type="checkbox"/>	Tug	Tuğba
<input type="checkbox"/>	Mek	Melike
<input type="checkbox"/>	Tan111	Taner
<input type="checkbox"/>	Sef1	Sefer
<input type="checkbox"/>	Vat2	Vater
<input type="checkbox"/>	Mut3	Mütter
<input type="checkbox"/>	Tah	Tahir
<input type="checkbox"/>	Bes11	Beste
<input type="checkbox"/>	Vat3	Alican

Aktion



Werkzeuge: EXAKT

Suche nach transkribierten oder annotierten Phänomenen

EXAKT

File Edit

RECENTLY USED

Corpora

- Dolmetschen im Krankenhaus**
S:\TP-Z...2\0,6\K2_Corpus.xml
111 transcriptions
20710 segment chains
- K5_Toerring**
S:\TP-Z...ing\K5_Toerring.xml
2 transcriptions
3484 segment chains
- K5_Oeresund**
S:\TP-Z...und\K5_Oeresund.xml
35 transcriptions
7364 segment chains

Done.

Concordances

- eigentlich**
Dolmetschen im Krankenhaus
79 tokens
1 types
- det**
K5_Toerring
846 tokens
1 types
- myck, mycket**
K5_Oeresund
208 tokens
2 types
- ver-me, dizia-me, fui-me, Anda-me**
Dolmetschen im Krankenhaus
152 tokens
106 types

Dolmetschen im Krankenhaus (79 results) | K5_Toerring (846 results) | K5_Oeresund (208 results) | Dolmetschen im Krankenhaus (152 results)

RegEx (T) Search: \b[A-Za-z]+me\b

S	C...	Speaker	Left Context	Match	Right Context	Type	Diskursar...	Konstellat...
✓ 91	SGa		E eu	meti-me	no carro, feito maluco, vim soz		Aufklärung	Monolingual
✓ 92	INa		mentos. ((2s)) ((schnalzt)) Eu,	falta-me	um exame, ((1,5s)) falta-me um	exclamative	Aufklärung	Monolingual
✓ 92	INa		Eu, falta-me um exame, ((1,5s))	falta-me	um exame para-lhe/ para res	other	Aufklärung	Monolingual
✓ 92	DZé		liguei muito. ((1s)) Ainda me ((deixe-me	andar). _E depois aquilo eh pas		Aufklärung	Monolingual
✓ 83	IFre		E eu quem sou?	Conhece...	a mim?	declarative	Anamnese	Monolingual
✓ 94	DMar		((2s)) Olhe,	deu-me	((unverständlich, 1s)) e a dar-		Anamnese	Monolingual
✓ 94	DMar		u-me ((unverständlich, 1s)) e a	dar-me	assim tonturas pela cabeça.		Anamnese	Monolingual
✓ 94	DMar		eu médico lá de família ace/ eh	receitou-me	•• aconselhou-me para fazer te	interrogative	Anamnese	Monolingual
✓ 94	DMar		família ace/ eh receitou-me ••	aconselho...	para fazer terapia e fui fazer	interrogative	Anamnese	Monolingual
✓ 94	DMar		ara fazer terapia e fui fazer e	encontrei...	muito mal.		Anamnese	Monolingual
✓ 94	DMar		á bastante tempo, mas o segundo	deu-me	cá de uma maneira. ((holt hörba		Anamnese	Monolingual
✓ 94	DMar		/ ao girar assim com o pescoço,	estalar-me	aqui muito. Foi por isso que o	declarative	Anamnese	Monolingual
✓ 94	DMar		nturas não. Deitada não, mas ao	por-me	a pé, vou já cair.		Anamnese	Monolingual
✓ 101	ILu		((blättern, 2 s)) Ora	deixe-me	cá ver. ((blättern, 3,5 s)) Ist	interrogative	Befund	Monolingual
✓ 101	ILu		Eu percebi,	estou-me	a rir, mas percebi.	interrogative	Befund	Monolingual
✓ 101	Umb		que me tinha acabado, não é, e	veio-me	e depois nunca mais me veio. Ac	exclamative	Befund	Monolingual
✓ 18	Mar		Ah, eu	sinto-me	bern.	declarative	Aufnahme...	Gedolmets...
✓ 18	Mar		Sim	Sinto-me	a calhar doente ((stoffert 10s	other	Aufnahme	Gedolmets...

Types: 106
Tokens: 152
Selected: 152
Time: 32.14 s

((blättern, 2 s)) Ora **deixe-me** cá ver. ((blättern, 3,5 s)) Isto hoje foi complicado. ((blättern, 2 s))

Type	interrogative
Diskursart[C]	Befund
Konstellationstyp[C]	Monolingual

Partitur

A [v] É? A senhora sabe algum? ((blättern, 2 s)) Ora **deixe-me** cá ver. ((blättern, 3,5 s)) Isto hoje foi complicado.

A [de] Ja? Wissen Sie welche? Also, lassen Sie mich sehen. Das heute war kompliziert.

P [v] É. Não é? É. Tá a ver?

P [de] Ja. Nicht wahr? Ja. Sehen Sie?

Partitur HTML

Gliederung

- Projektrahmen und Projekt
- EXMARaLDA Architektur und Werkzeuge
- Drei SFB-Korpora**
- Demonstration „Skandinavische Semikommunikation“
- Aufarbeitung der Korpora
- Korpus-Fertigstellung
- Neue Möglichkeiten -> neue Probleme
- Anleitung zum Arbeit-sparen
- Plädoyer für bessere Datenarbeit

Drei Korpora gesprochener Sprache

Dolmetschen im Krankenhaus

Skandinavische Semikommunikation

**Sprachliche Konnektivität bei bilingual türkisch-deutsch
aufwachsenden Kindern**

Drei Korpora gesprochener Sprache

Dolmetschen im Krankenhaus

25h Audiomaterial, 112 Transkriptionen, ca. 170.000 Wörter

Skandinavische Semikommunikation

90h Audiomaterial, ca. 50% transkribiert in
74 Transkriptionen, ca 300.000 Wörter

Sprachliche Konnektivität bei bilingual türkisch-deutsch aufwachsenden Kindern

ca. 700 Transkriptionen und 700.000 Wörter

**Alle Korpora liegen nach HIAT –Konventionen transkribiert
in EXMARaLDA vor.**

Gliederung

- Projektrahmen und Projekt
- EXMARaLDA Architektur und Werkzeuge
- Drei SFB-Korpora
- Demonstration „Skandinavische Semikommunikation“**
- Aufarbeitung der Korpora
- Korpus-Fertigstellung
- Neue Möglichkeiten -> neue Probleme
- Anleitung zum Arbeit-sparen
- Plädoyer für bessere Datenarbeit

Skandinavische Semikommunikation

6 Subkorpora

- NUAS („Nordic Association of University Administrators“) / Aufgenommene Konferenzgespräche
- Öresund Direkt (Radioaufnahmen)
- Radio-Recordings (Radioaufnahmen)
- Deutsche Schule (Schulstunden)
- Dänische Schule (Schulstunden)
- Universitätskurse

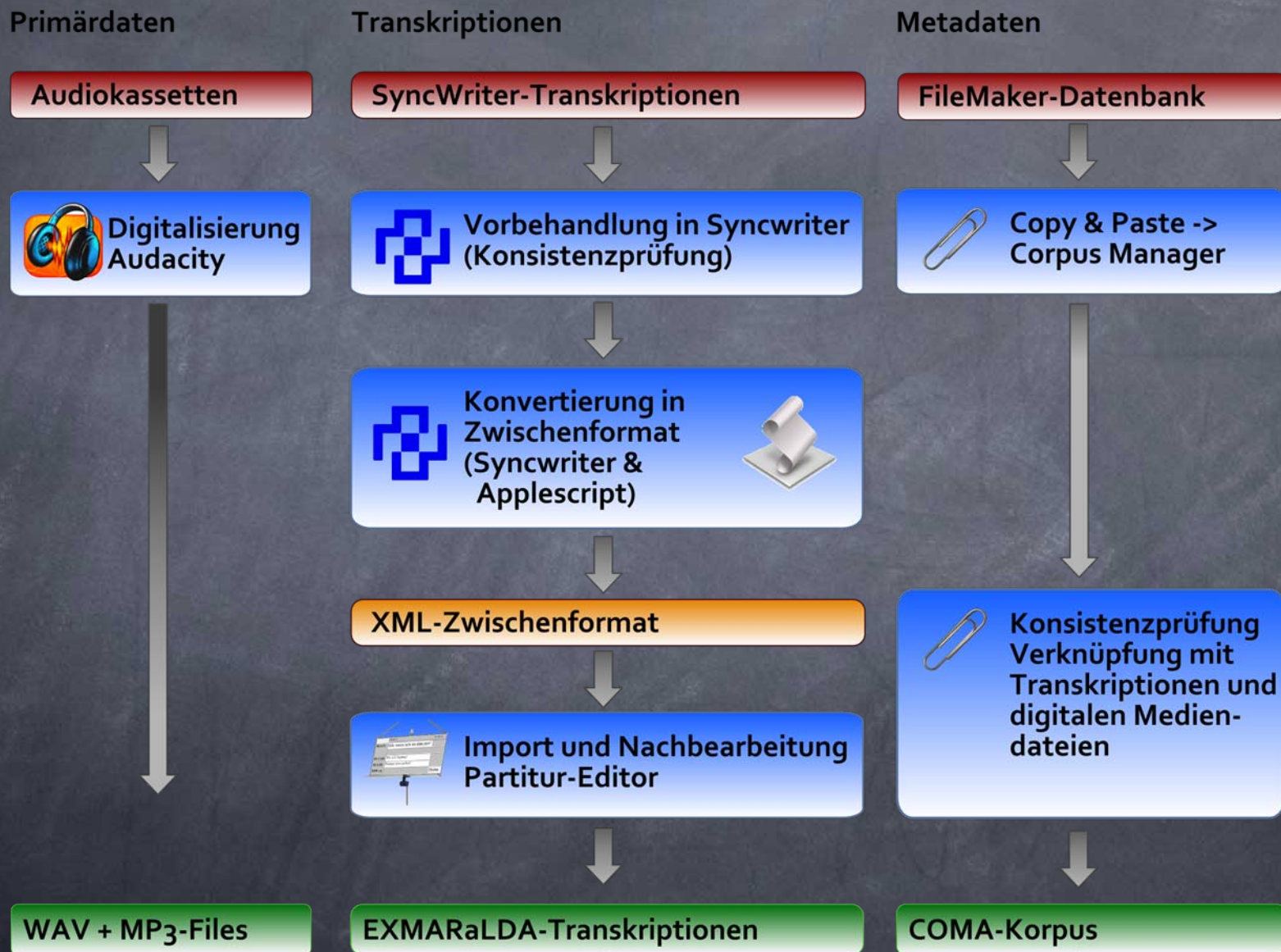
Demo Skandinavische Semikommunikation



Gliederung

- Projektrahmen und Projekt
- EXMARaLDA Architektur und Werkzeuge
- Drei SFB-Korpora
- Demonstration „Skandinavische Semikommunikation“
- Aufarbeitung der Korpora**
- Korpus-Fertigstellung
- Neue Möglichkeiten -> neue Probleme
- Anleitung zum Arbeit-sparen
- Plädoyer für bessere Datenarbeit

Aufarbeitung der Korpora: SKOBI/ENDFAS



Aufarbeitung der Korpora

Skandinavische Semikommunikation

- Konvertierung aus Praat quasi automatisch
- Metadaten nicht digital oder gar nicht erfasst
- -> Ermittlung und Eingabe der Metadaten

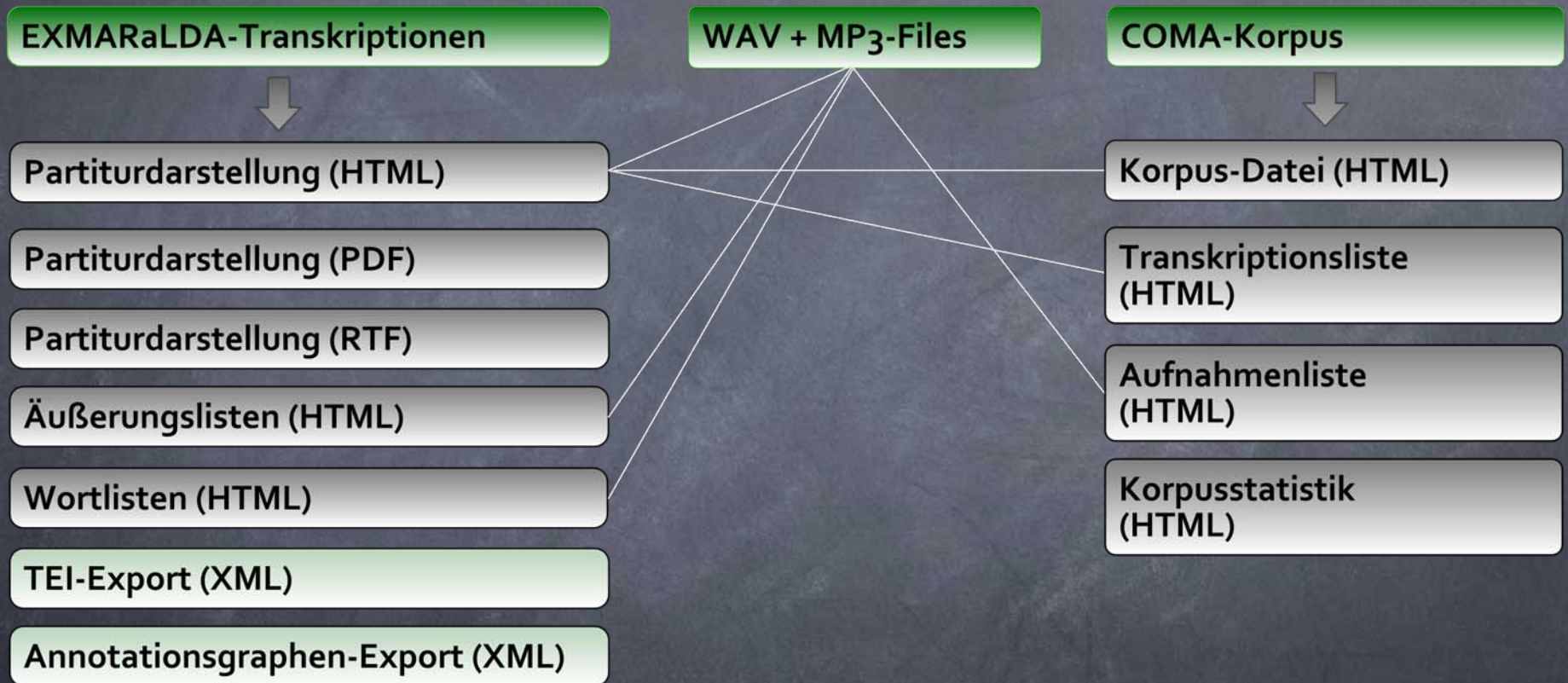
Dolmetschen im Krankenhaus

- Konvertierung SyncWriter -> EXMARaLDA im Projekt
- Nachbearbeitung und Konvertierung von Metadaten

Gliederung

- Projektrahmen und Projekt
- EXMARaLDA Architektur und Werkzeuge
- Drei SFB-Korpora
- Demonstration „Skandinavische Semikommunikation“
- Aufarbeitung der Korpora
- Korpus-Fertigstellung**
- Neue Möglichkeiten -> neue Probleme
- Anleitung zum Arbeit-sparen
- Plädoyer für bessere Datenarbeit

Korpus-Fertigstellung



Angepasste XSL-Stylesheets und Batch-Dateien

Gliederung

- Projektrahmen und Projekt
- EXMARaLDA Architektur und Werkzeuge
- Drei SFB-Korpora
- Demonstration „Skandinavische Semikommunikation“
- Aufarbeitung der Korpora
- Korpus-Fertigstellung
- Neue Möglichkeiten -> neue Probleme**
- Anleitung zum Arbeit-sparen
- Plädoyer für bessere Datenarbeit

Neue Möglichkeiten -> Neue Probleme

Neue Möglichkeiten: „Corpus-Driven Linguistics“

Neue Probleme: Werkzeuge decken Fehler auf

Gliederung

- Projektrahmen und Projekt
- EXMARaLDA Architektur und Werkzeuge
- Drei SFB-Korpora
- Demonstration „Skandinavische Semikommunikation“
- Aufarbeitung der Korpora
- Korpus-Fertigstellung
- Neue Möglichkeiten -> neue Probleme
- Anleitung zum Arbeit-sparen**
- Plädoyer für bessere Datenarbeit

Anleitung zum Arbeit-sparen

Keine Software, die proprietäre, undokumentierte Formate benutzt!

Anleitung zum Arbeit-sparen

Keine Software, die proprietäre, undokumentierte Formate benutzt!

Metadaten erheben. Metadaten elektronisch erheben.

Anleitung zum Arbeit-sparen

Keine Software, die proprietäre, undokumentierte Formate benutzt!

Metadaten erheben. Metadaten elektronisch erheben.

Dokumentieren!

Metadatenerhebung, Transkriptionskonventionen, Bearbeiter und Bearbeitungsschritte, Entscheidungen für Datenformate, ...

Gliederung

- Projektrahmen und Projekt
- EXMARaLDA Architektur und Werkzeuge
- Drei SFB-Korpora
- Demonstration „Skandinavische Semikommunikation“
- Aufarbeitung der Korpora
- Korpus-Fertigstellung
- Neue Möglichkeiten -> neue Probleme
- Anleitung zum Arbeit-sparen
- Plädoyer für bessere Datenarbeit**

Plädoyer für bessere Datenarbeit

Gedanken vorher machen

Was soll das Korpus können? Wer soll damit arbeiten können?

Plädoyer für bessere Datenarbeit

Gedanken vorher machen

Was soll das Korpus können? Wer soll damit arbeiten können?

Noch Leichen im Keller?

Audi-Videodaten **SOFORT** digitalisieren!

Metadaten **SOFORT** erfassen!

Durchatmen



Plädoyer für bessere Datenarbeit

Gedanken vorher machen

Was soll das Korpus können? Wer soll damit arbeiten können?

Noch Leichen im Keller?

Audi-Videodaten **SOFORT** digitalisieren!

Metadaten **SOFORT** erfassen!

Durchatmen

Lassen sich die Transkriptionen retten?

Danke

Deutsche
Forschungsgemeinschaft

DFG



Universität Hamburg

Sonderforschungsbereich
Mehrsprachigkeit

