Automatische Zielhypothesen in DAKODA

Josef Ruppenhofer, Torsten Zesch

September 25, 2025

DAKODA

Outline

DAKODA

Zielhypothesen

- Zielhypothesen in der Lernerkorpuslinguistik verwendet, um eine erklärbare und explizite Fehlerannotation zu ermöglichen [Lüdeling, 2008, Reznicek et al., 2013]
- Beispiel aus Falko-Whig [Krummes and Ensslin, 2014]:

Base text | Token Annotations

1 / 1 Displaying Results 1 - 1 of 1 Result for: "gehte"

Path: FalkoWHIGL2v2.1 > BNG-2010-04-074 (tokens 106 - 116) left context: 5 right context: 5

verlassen	mussten	.	Der	Mann	gehte	jeden	Tag	zur		Arbeit	,
verlassen	müssen	.	d	Mann	[unknown]	jed	Tag	zur		Arbeit	,
VVINF	VMFIN	$.	ART	NN	VVFIN	PIAT	NN	APPRART		NN	$,

⊟ ZH0 (grid)

ZH0	verlassen	mussten	.	Der	Mann	ging	jeden	Tag	zur	Arbeit	,
ZH0Diff						CHA					
ZH0S	s5			s6							
ZH0lemma	verlassen	müssen	.	d	Mann	gehen	jed	Tag	zur	Arbeit	,
ZH0lemmaDiff						CHA					
ZH0pos	VVINF	VMFIN	$.	ART	NN	VVFIN	PIAT	NN	APPRART	NN	$,
ZH0rfMorph	Full2	Mod.3.Pl.Past.Ind	Pun.Sent	Def.Nom.Sg.Masc	Reg.Nom.Sg.Masc	Full.3.Sg.Past.Ind	Indef.Attr3.Acc.Sg.Masc	Reg.Acc.Sg.Masc	Dat.Sg.Fem	Reg.Dat.Sg.Fem	Pun.Comma
ZH0rfPos	VVINF	VMFIN	$.	ART	NN	VVFIN	PIAT	NN	APPR	NN	$,
tok	verlassen	mussten	.	Der	Mann	gehte	jeden	Tag	zur	Arbeit	,

Figure 1: "Der Mann ~~gehte~~ ging jeden Tag zur Arbeit"

DAKODA

Zielhypothesen II

- Verschiedene Typen
- Falko-Korpus-Familie [Reznicek et al., 2012]
  - **ZH1**: morpho-syntaktische Korrekturen
  - ZH2: auch semantisch-pragmatische Verbesserungen
  - ZH0: wie ZH1, aber ohne Veränderungen der Verbstellung
- orthografische ZHs [Laarmann-Quante et al., 2017]
- Intermediate + final THs in Dulko [Beeh et al., 2021]

DAKODA

Zielhypothesen II

- zweckgebunden, keine Einheitslösung
- L1 als Referenz / Norm (pace Bley-Vroman [1983])
- fehlerorientiert
- explizit
- überwiegend satzorientiert: keine Umformulierung des Textes

DAKODA

Zielhypothesen in Dakoda

- Allgemein: für die Community zur Ermöglichung von Fehlerannotation/-untersuchung

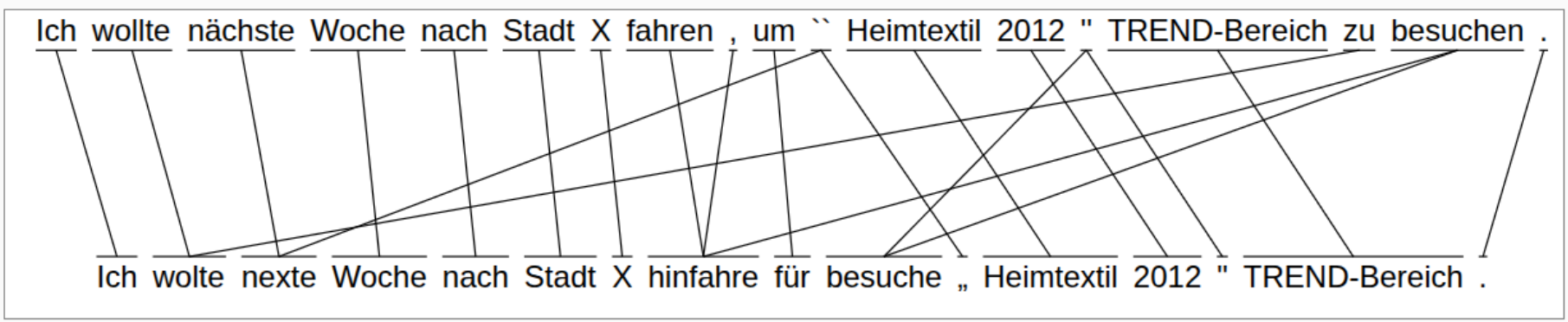
Zielhypothesen in Dakoda

- Allgemein: für die Community zur Ermöglichung von Fehlerannotation/-untersuchung
- Als Normalisierungslayer zur Verbesserung der Zugänglichkeit von Lernerdaten (z.B. orthographische Normalisierung)

DAKODA

Alignment zwischen ZHs und Lernerspur

- Viele Alignment-Varianten verfügbar, keine perfekt
- DAKODA verwendet *simalign* [Jalili Sabet et al., 2020] (MWMF)
- Post-processing um z.B. Mehrfach-Alignments zu eliminieren



- Sentence alignment mit *SentAlign* [Steingrimsson et al., 2023] vorgeschaltet.

Darstellung https://vilda.net/s/slowalign/

DAKODA

Vergleich von Samples mehrerer automatischer und manueller ZHs

AB(CD)-Test

	A2 sentence	A1 sentence	
1	Ich möcht trefen mit dir .		Lerner
2	Ich möchte mich mit dir treffen .	.	
3	Ich möchte mich treffen mit dir .		
4	Ich möchte mich treffen mit dir .		
5	Ich möchte dich treffen .		

Vergleich von Samples mehrerer automatischer und manueller ZHs

AB(CD)-Test

	A2 sentence	A1 sentence	
1	Ich möcht trefen mit dir .	Leider können Sie aber nicht Fahren Wichtig ich muss am Wochenende arbeiten .	Lerner
2	Ich möchte mich mit dir treffen .	Leider können Sie aber nicht fahren , wichtig , ich muss am Wochenende arbeiten .	
3	Ich möchte mich treffen mit dir .	Leider können Sie aber nicht fahren . Wichtig : Ich muss am Wochenende arbeiten .	
4	Ich möchte mich treffen mit dir .	Leider können Sie aber nicht fahren . Wichtig : Ich muss am Wochenende arbeiten .	
5	Ich möchte dich treffen .	Leider können Sie aber nicht fahren . Wichtig ist : Ich muss am Wochenende arbeiten .	

DAKODA

Vergleich von Samples mehrerer automatischer und manueller ZHs

AB(CD)-Test

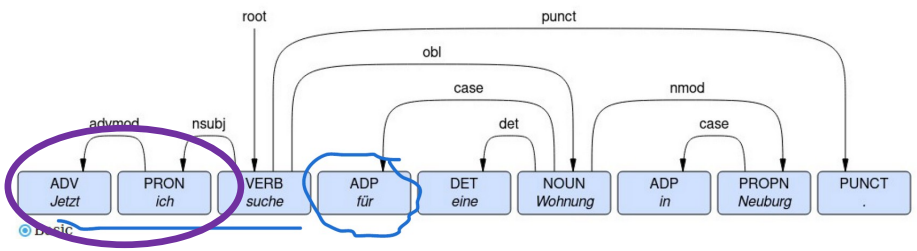
	A2 sentence	A1 sentence	
1	Ich möcht trefen mit dir .	Leider können Sie aber nicht Fahren Wichtig ich muss am Wochenende arbeiten .	Lerner
2	Ich möchte mich mit dir treffen .	Leider können Sie aber nicht fahren , wichtig , ich muss am Wochenende arbeiten .	transgec
3	Ich möchte mich treffen mit dir .	Leider können Sie aber nicht fahren . Wichtig : Ich muss am Wochenende arbeiten .	manual
4	Ich möchte mich treffen mit dir .	Leider können Sie aber nicht fahren . Wichtig : Ich muss am Wochenende arbeiten .	mbartgec
5	Ich möchte dich treffen .	Leider können Sie aber nicht fahren . Wichtig ist : Ich muss am Wochenende arbeiten .	mixtral

Probleme beim Parsen von Lernerdaten

L1 Bias

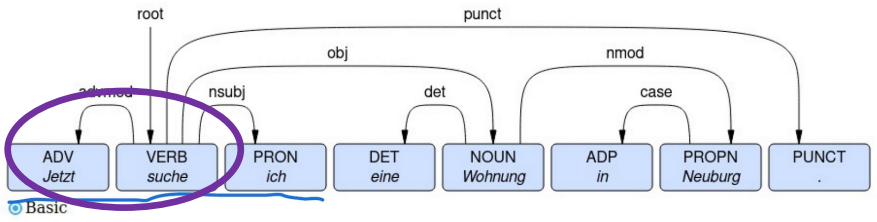
Beispiel: parallele Parses für Lernerspur und ZH

learner



SVO

ATH1 (mixtral)



INV

DAKODA

Probleme beim Parsen von Lernerdaten

Schmetterlingseffekte

L: Dies ist eine umstrittene Frage , [den um Heute überhaupt einen Job zu kriegen , wird **verlangt** dass man einen Universitätsabschluss hat].

ZH1: Dies ist eine umstrittene Frage , [denn um heute überhaupt einen Job zu kriegen , wird **verlangt** , dass man einen Universitätsabschluss hat]

DAKODA

Parsen von Lernerdaten

Beispiel spaCy

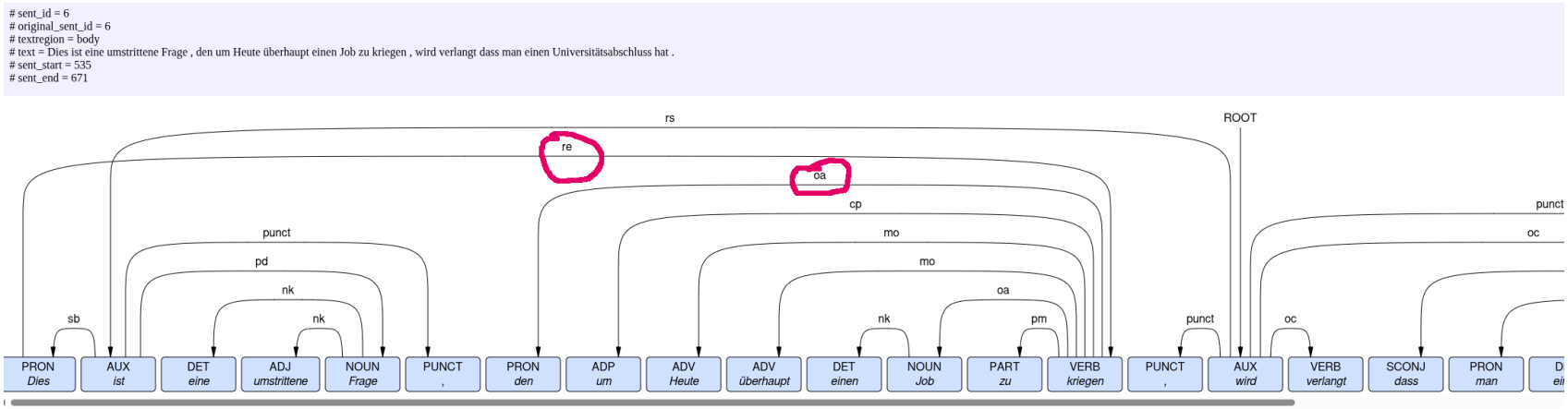


Figure 2: spaCy-Parse

- *den* als Demonstrativpronomen und Objekt (unmögliche Valenz von *kriegen*)
- Gleichzeitig *Heute* korrekt als Adverb behandelt
- Adverbialsatz fälschlicherweise an das Subjekt des linken Konjunkts angehängt (RE)
- Vorfeld des rechten Konjunkts nicht Adverbialsatz sondern das linke Konjunkt (RS)
- 💥 *verlangt* hat kein Subjekt

DAKODA

Parsen von Lernerdaten

Beispiel syntaxdot

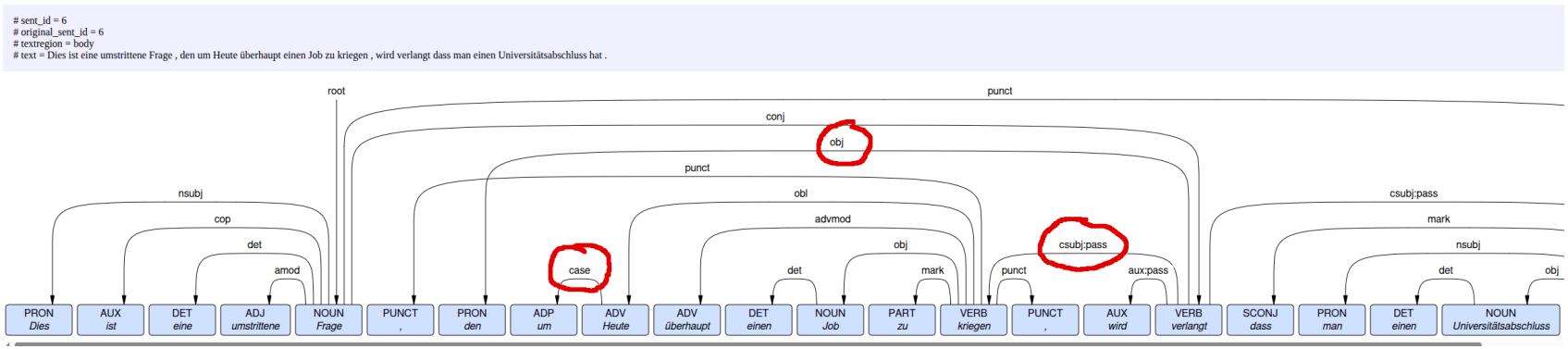


Figure 3: syntaxdot-Parse

- *den* als Objekt von *kriegen*; *Heute* als Adverb
- Adverbialsatz hier Subjektsatz
- 💥 Verb im rechten Konjunkt hat zwei Subjekte
- Koordination der Teilsätze hier richtig

DAKODA

Parsen der ZH

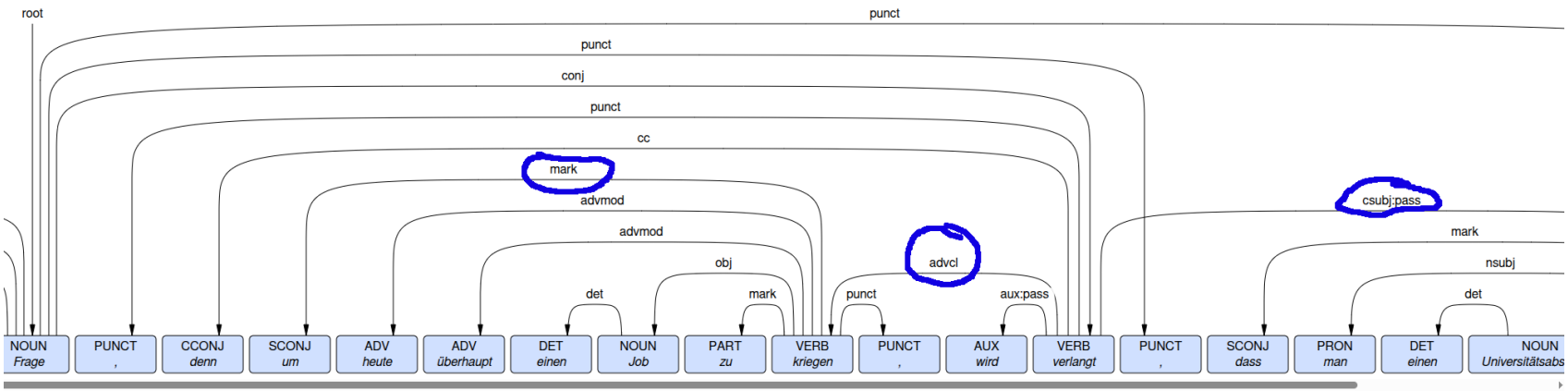


Figure 4: syntaxdot-Parse

- Alle Problemfälle der Lernerspur sind hier richtig analysiert: wir würden für das rechte Konjunkt korrekterweise INV(ersion) beobachten.

DAKODA

Zielhypothesen als Hilfsmittel beim Parsen der Lernerspur

- Aus linguistischer Sicht (DaF/DaZ) können wir Stufenanalysen auf einer ZH nicht anstelle von Stufenanalysen auf der Lernerspur verwenden:
  - ZHs (zumindest ZH1, ZH2) sollen grammatisch sein
  - ZHs müssen daher z.B. (pace Müller 2003) alle ADV-Instanzen eliminieren
  - flüssigere ZHs haben z.T. andere Linearisierungen: z.B. SEP-Einführung oder -Eliminierung durch Wechsel zwischen Präteritum und Perfekt
- (Aus praktischer CL/NLP-Sicht können wir aber trotzdem untersuchen, wie ähnlich die Ergebnisse wären.)

DAKODA

Zielhypothesen als Hilfsmittel beim Parsen der Lernerspur

- Idee: TH Parses als Korrektiv für Lernerspurparses vor der Stufenanalyse (vgl. Rehbein et al. [2012])

Zielhypothesen als Hilfsmittel beim Parsen der Lernerspur

- Idee: TH Parses als Korrektiv für Lernerspurparses vor der Stufenanalyse (vgl. Rehbein et al. [2012])
  - Wie genau?

DAKODA

Zielhypothesen als Hilfsmittel beim Parsen der Lernerspur

- Idee: TH Parses als Korrektiv für Lernerspurparses vor der Stufenanalyse (vgl. Rehbein et al. [2012])
  - Wie genau?
- Voraussetzung: Automatisierung der TH-Generierung

DAKODA

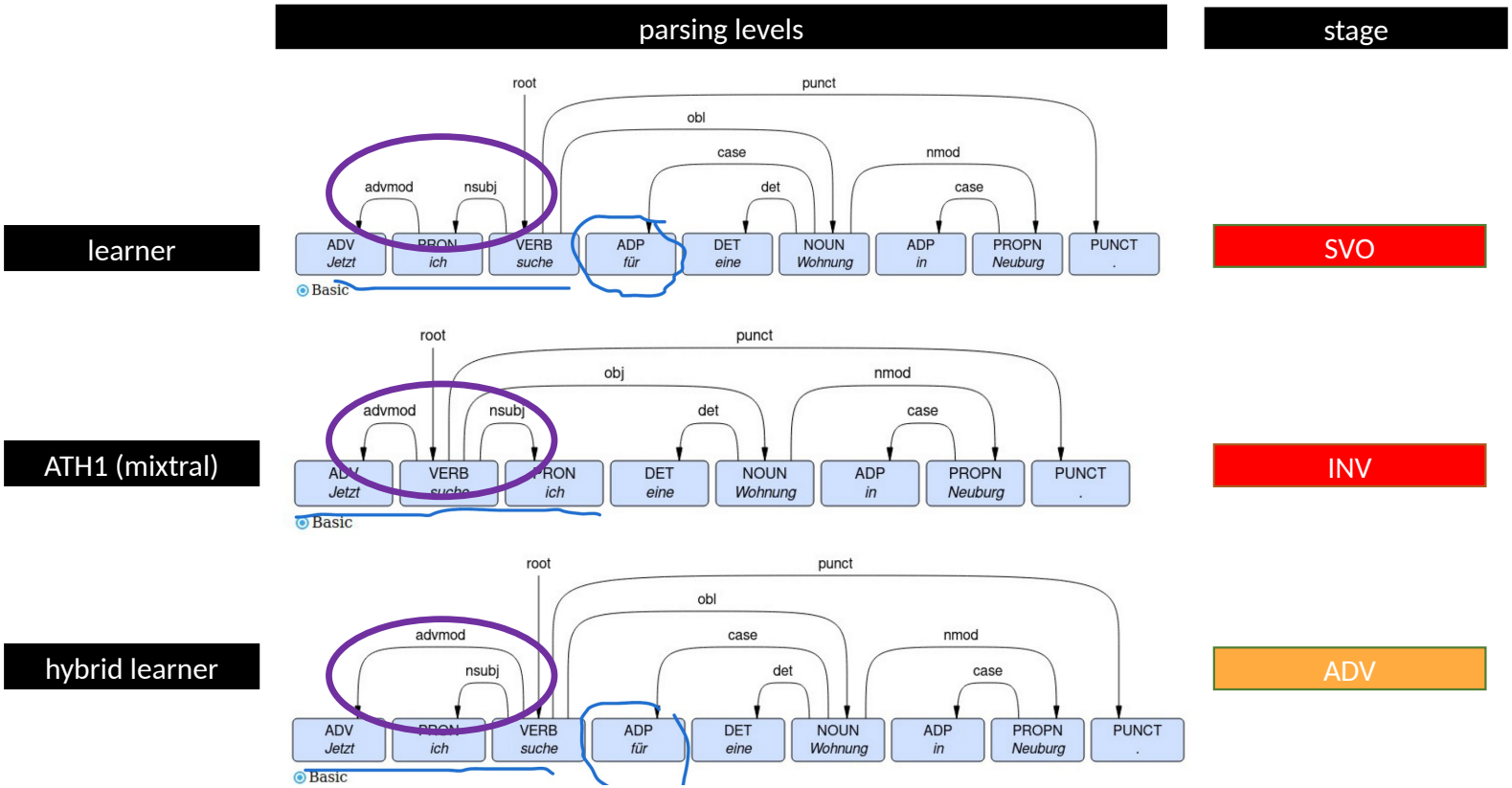
Zielhypothesen als Hilfsmittel beim Parsen der Lernerspur

- Idee: TH Parses als Korrektiv für Lernerspurparses vor der Stufenanalyse (vgl. Rehbein et al. [2012])
  - Wie genau?
- Voraussetzung: Automatisierung der TH-Generierung
  - Wie vergleichen sich automatische ZHs mit manuell erstellten?

DAKODA

Zielhypothesen als Hilfsmittel beim Parsen der Lernerspur

Hybridisierung



Tokens der Lernerspur (samt Linearisierung) beibehalten, aber ling. Interpretation anpassen -> Stufenzuordnung

DAKODA

Stufenzuweisung auf Lernerdaten

- Zwei Lerner-Datensätze
  - Kalibrierung: stratifiziertes Sample von Merlin + Disko-Daten
  - Six corpus sample
    - Bematac L2
    - CDPS
    - Kolipsi 1 (KLP1)
    - Multilit spoken (MULT_S)
    - Multilit written (MULT_W)
    - SWIKO/WTLD
- Ein L1(-artiger) Vergleichsdatensatz
  - UDTopoADV
    - Sätze aus UD-Baumbanken (HDT, GSD)
    - topologische Testbatterie
    - künstlich erzeugte ADV-Sätze (Basis: UD-Treebanks)

DAKODA

Manuelle Regeln

UDTopoADV

label	precision	recall	f1-score	support
ADV	0.00	0.00	0.00	65
INV	0.79	0.90	0.84	234
SEP	0.76	0.97	0.86	177
SVO	0.73	0.96	0.83	254
VEND	0.87	0.99	0.93	191
———	———	———	———	———
micro avg	0.78	0.88	0.83	921
macro avg	0.63	0.76	**0.69**	921
weighted avg	0.73	0.88	0.80	921

DAKODA

Manuelle Regeln

Kalibrierungssample

Label	precision	recall	f1-score	support
ADV	0.00	0.00	0.00	34
INV	0.97	0.90	0.93	470
SEP	0.83	0.85	0.84	327
SVO	0.95	0.89	0.92	537
VEND	0.92	0.99	0.95	418
———	———	———	———	———
micro avg	0.93	0.89	0.91	1786
macro avg	0.73	0.73	**0.73**	1786
weighted avg	0.91	0.89	0.90	1786

Manuelle Regeln

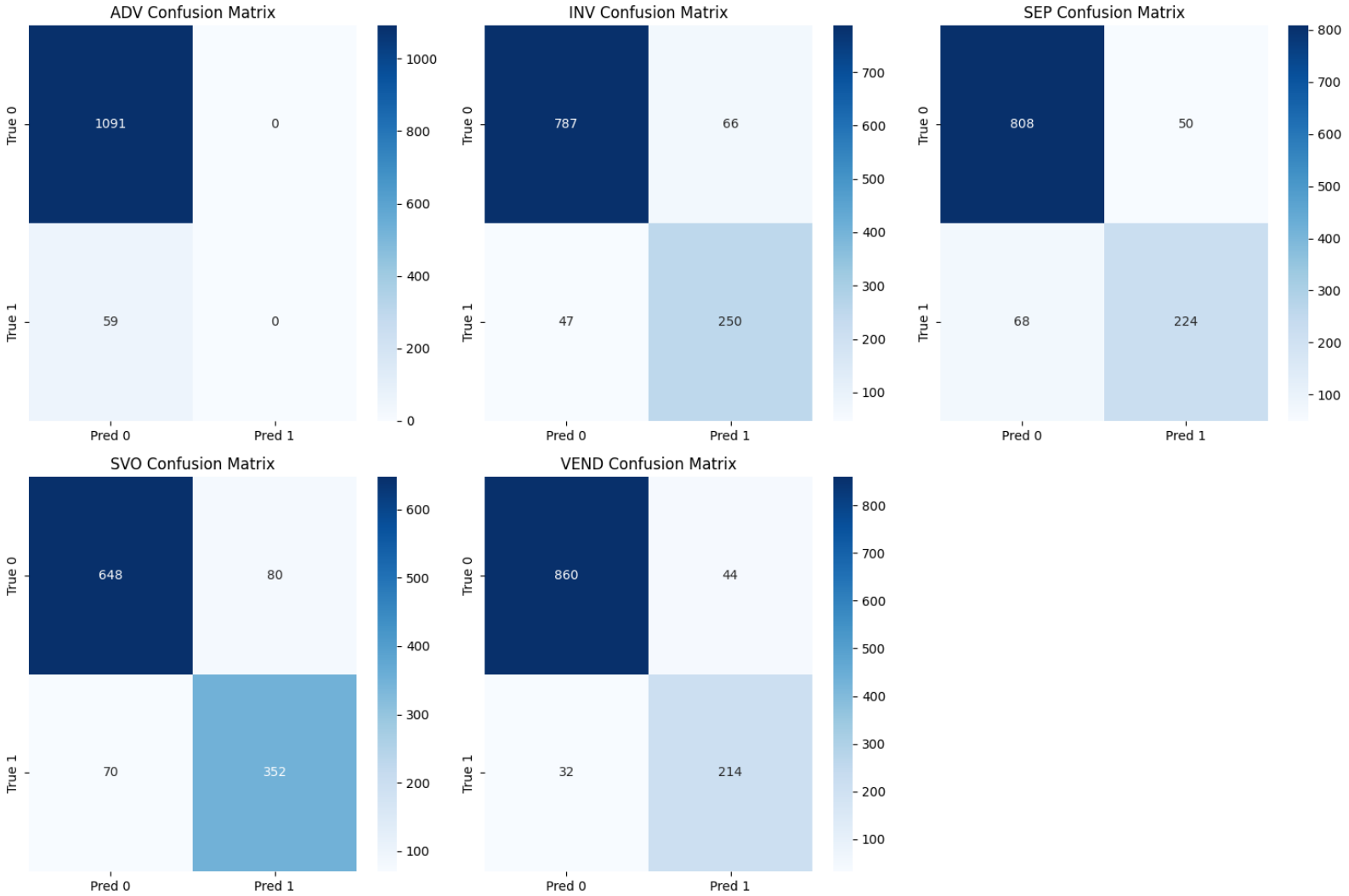
Six corpus sample

label	precision	recall	f1-score	support
ADV	0.00	0.00	0.00	59
INV	0.79	0.84	0.82	297
SEP	0.82	0.77	0.79	292
SVO	0.81	0.83	0.82	422
VEND	0.83	0.87	0.85	246
———	———	———	———	———
micro avg	0.81	0.79	0.80	1316
macro avg	0.65	0.66	**0.66**	1316
weighted avg	0.78	0.79	0.78	1316

DAKODA

Manuelle Regeln

Konfusionsmatrix für Six corpus sample



Stufenvorhersage für Lernerdaten

Zwischenfazit

- Für L1-kompatible Stellungen im Großen und Ganzen gute Ergebnisse
- Bei ADV aber nahezu keine Erkennung.
- Was tun?

DAKODA

Wissenstransfer von ZH zu Lernerspur

Kontexte für Transfer

- Selektiver, durch bestimmmte Dependenzrelationen ausgelöste Übertragung (z. B. vocative, discourse)
- Selektiver Transfer für bestimmte Wortarten wie ADV, INTJ, PART
- Vollständiger Transfer, wenn alle Nicht-Interpunktions-Token eines ganzen Satzes aligniert sind
- Wenn das finite Verb bzw. die Verbklammer eines (Teil)Satzes sowie deren Kinder über die Lerner- und TH-Spuren hinweg aligniert sind, übertragen wir ggf. POS- und Dependenzinfo für diese Tokens (lassen aber deren Kinder unverändert)
- Alles unter dem Vorbehalt, dass ein wohlgeformter Baum erhalten bleibt.

DAKODA

Manuelle Regeln auf rohen vs hybridisierten Lernerparses

Kolipsi-1 L2 sample

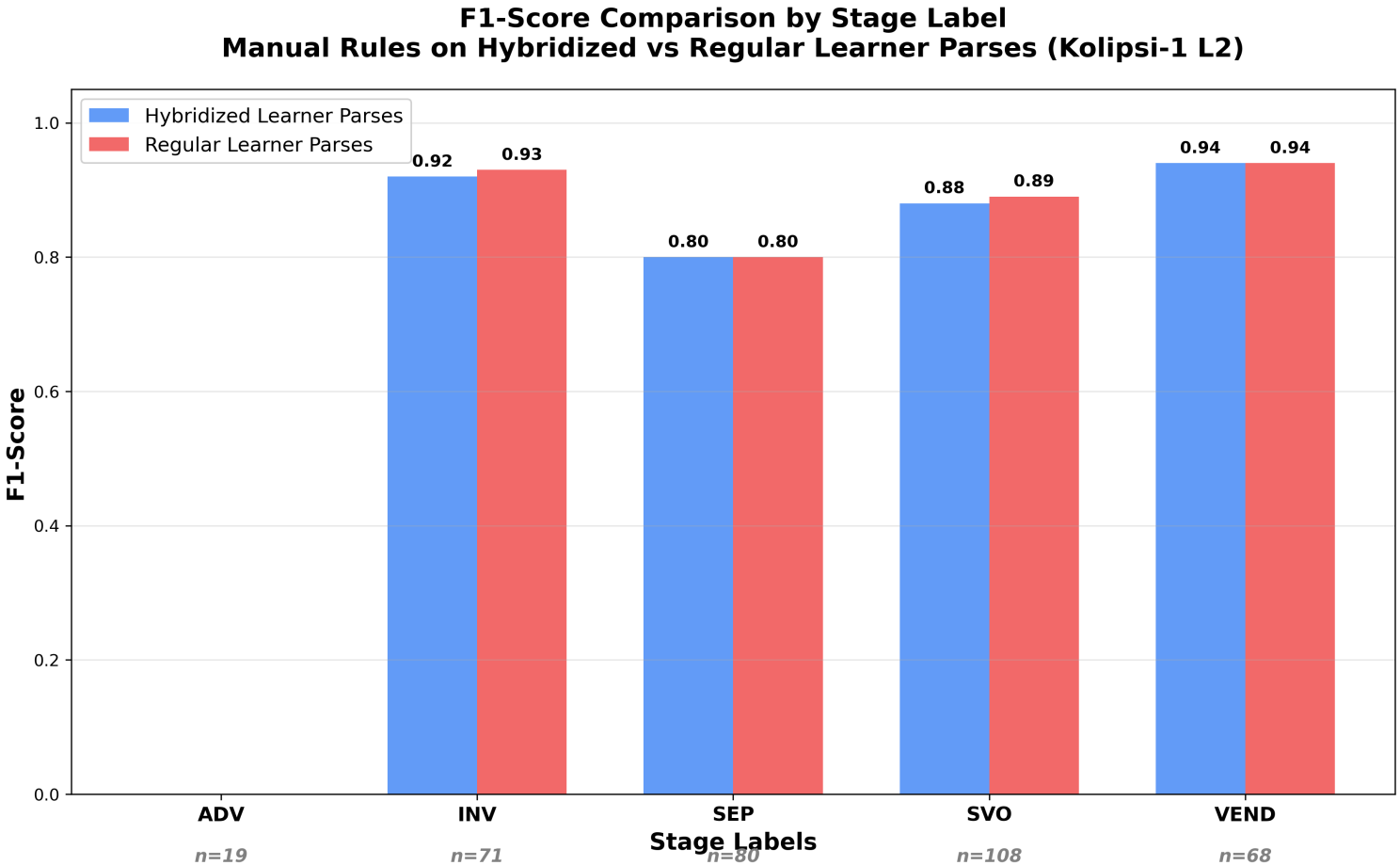
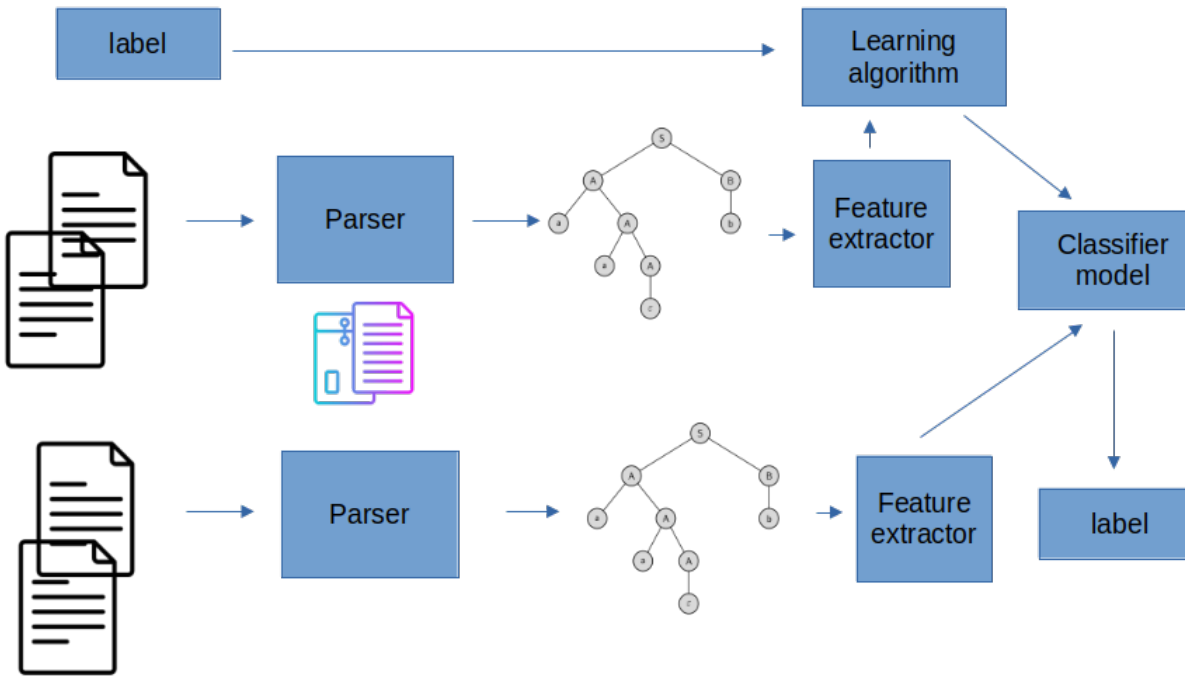


Figure 1: F1-score comparison across labels

DAKODA

Verbesserung der Ergebnisse

Weitere Optionen



DAKODA

Verbesserung der Ergebnisse

Weitere Optionen

- Konsistentere Labels? eigentlich gutes Agreement [Ruppenhofer et al., 2025]
- Besseren / robusteren L1-Parser suchen (für unsere Zwecke) ✅ (siehe Anhang; gestestet als Input für Stufenvorhersagen auf dem Six corpus sample)
- Parser für L2 adaptieren (sprengt Projektrahmen ❌)
- Parserkombination: Wir amalgamieren mehrere Parses zu einem als Input für die Stufenerkennung [Attardi and Dell'Orletta, 2009, Barry et al., 2020] (siehe Anhang; leider keine Verbesserung 💔 😐)
- Wir lassen die Stufenerkennung auf parallelen Parses laufen und kombinieren hinterher die zugewiesenen Labels. 😐
- Bessere Klassifikatoren und Ensemble-Systeme ✅ (siehe Anhang; vorläufig getestet im in-domain setting 💚)
- Wir importieren Wissen aus parallelen / verwandten Textversionen (leider bisher keine Verbesserung 😐)
- Mehr Instanzen (zumindest für ADV)

DAKODA

Erstellung von Zielhypothesen in Dakoda

- Automatisch

DAKODA

Erstellung von Zielhypothesen in Dakoda

- Automatisch
- zwei unserer drei Systeme sind automatische Grammatical Error Correction Systeme, die auf Falko+Merlin (Boyd et al. 2014) trainiert sind

DAKODA

Erstellung von Zielhypothesen in Dakoda

- Automatisch
- zwei unserer drei Systeme sind automatische Grammatical Error Correction Systeme, die auf Falko+Merlin (Boyd et al. 2014) trainiert sind
- LLM als drittes System

Erstellung von Zielhypothesen in Dakoda

- Automatisch
- zwei unserer drei Systeme sind automatische Grammatical Error Correction Systeme, die auf Falko+Merlin (Boyd et al. 2014) trainiert sind
- LLM als drittes System
- Unterschiede zwischen ZH-Systemen erlaubt/erwünscht

DAKODA

Mehrere Zielhypothesen: Mehrere Normalisierungen

	Wann möchst wir Treffen ?	**Dein Babys ist Name ?**
Gloss	when want.2SG we.1PL.Nom meeting ?	your.SG baby.{PL,SG} is name.SG
HR0	Wann möchten wir uns treffen?	Dein Baby hat einen Namen?
HR1	Wann möchtest du , dass wir uns treffen ?	Wie heißen deine Babys ?
HR2	Wann möchtest du dich treffen ?	Wie heißt dein Baby ?
HR3	Wann möchtest du dich mit mir treffen ?	Wie ist der Name deines Babys ?
HR4	Wann möchtest du ein Treffen?	Was ist der Name deines Babys ?

DAKODA

Mehrere Zielhypothesen: Unklare Bedeutung / Äußerungsabsicht

Learner [Ich frage meinen Freundin Julia brauchen tiket für konzert]. **[und Sie sagen mir gut Konzert wurde 18 märz]**. [und ich zürük fahren 28 märz] (Source: Merlin 1091_0000254)

TH1 [Ich frage meine Freundin Julia , ob sie ein Ticket für ein Konzert braucht] **Und Sie sagen mir , ob das Konzert am 18 . März gut wurde** . [Und ich fahre am 28 . März zurück].

TH2 [Ich frage meine Freundin Julia , ob sie ein Ticket für das Konzert braucht]. **Sie wird mir zusagen . Das Konzert wäre am 18 . März** [und ich komme am 28 . März zurück]

DAKODA

Zielhypothesen: untersuchte Systeme

- mbartgec - mbart-german-grammar-corrector
- transgec - TransGEC: Improving Grammaticial Error Correction with Translationese [Fang et al., 2023]
- mixtral LLM
- weitere, nicht in der Dakoda-Daten-Release enthaltene
  - llama3 LLM [Touvron et al., 2023]
  - hessianai leo 13b LLM
  - "Roundtrip" Übersetzung: $\text{DE}^{L2} => \text{EN} => \text{DE}^{L1}$ (mbart)
  - Direkte Übersetzung: $\text{DE}^{L2} => \text{DE}^{L1}$ (mbart)

DAKODA

Gemeinsamkeiten

- GEC / Fehlerkorrektur konzeptualisiert wie folgt: Eingangstextsequenz in Ausgagbetextsequenz übersetzen ("Sequence-to-sequence").
- Gängige Ansätze benutzen in Teilen oder in Gänze die sog. Transformer-Architektur:
  - Encoder: Verarbeitet die Eingabesequenz.
  - Decoder: Erzeugt die Ausgabesequenz.
  - Cross-Attention: Verbindet Encoder und Decoder.
- LLMs verwenden nur den Decoder: Vervollständigung / Fortsetzen von Text
- Bei Transgec kommen z.T. ganze Transformer zum Einsatz; mbart(gec) basiert auch auf ganzen Transformern.
- (BERT benutzt nur den Encoder, üblicherweise nicht für Textgenerierung verwendet)

DAKODA

Mbartgec

- Grundsystem Multilingual Bart (mBart) [Liu et al., 2020] ist ein *multilinguales* MÜ-System, das vor dem Training für die eigentliche Übersetzungsaufgabe auf einer sog. 'denoising'-Task trainiert wurde.
- Denoising-Task: künstlich verfremdete (korrumpierte) Texte wieder korrigieren bzw. Originaltext wieder herstellen
  - Tokens des Originaltexts maskiert
  - Permutierung von Sätzen
- Fine-tuning von mBart durch mbartgec: mbartgec trainiert das Modell auf deutschen Fehlerkorrektur-Daten (Falko-MERLIN Datensatz für Grammatical Error Correction [Boyd, 2018]) weiter.
  - 'Fine-getuntes' Modell nutzt vorhandenes 'Allgemeinwissen', erlernt weitere aufgabenspezifische Fähigkeiten.
  - Braucht weniger Trainingsbeispiele als das Trainieren eines Fehlerkorrekturmodells von Grund auf.

DAKODA

Transgec

- Übliche Ansätze zur "Datenerweiterung" für GEC (Fehlerkorrektur) verfremden L1-Texte.
- Diskrepanz: eigentliche Aufgabe besteht darin, Texte zu korrigieren, die von Nicht-Muttersprachlern verfasst wurden, während für das Training künstlich verfälschte Texte von Muttersprachlern verwendet werden, die einen anderen "Stil" aufweisen.
- Idee: lieber Translationese verwenden.
- 'Translationese': Texte, die ursprünglich in einer Sprache verfasst und dann von Menschen in eine andere Sprache übersetzt wurden.
- Annahme: Translationese oft etwas „seltsam", ähnlich wie Texte, die nicht von Muttersprachlern verfasst wurden: umständliche Formulierungen, ungewöhnliche Wortwahl , grammatikalische Muster, die von der L1 beeinflusst sind.
- Translationese wird zusätzlich verfremdet: Löschungen, Einfügungen, Ersetzungen, Umordnung.
- In Kombination mit einem GEC-Grundkorpus (z.B. Merlin-Falko von A. Boyd) wird dann auf den verfremdeten Daten ein Übersetzungsmodell (weiter)trainiert.

DAKODA

Mixtral

- LLMs lernern vorherzusagen, welches Wort oder welcher Ausdruck als nächstes in einer Sequenz kommen sollte.
- Beispiel: „Die Katze saß auf dem ___“ „Teppich“, „Stuhl“ oder „Dach“
- ‘Large’ language models lernen auf riesigen Datenmengen.
- Durch diesen Prozess entwickeln sie ein Verständnis für Grammatik, ‘Fakten’, Schreibstile.
- Mixtral verwendet Ansatz namens „Mixture of Experts” (MoE), um sowohl leistungsstark als auch effizient zu sein.
- Anstelle eines einzigen Netzwerks, das alles verarbeitet, hat Mixtral mehrere kleinere Spezialisten-Netzwerke.
- Ein “Dispatcher” entscheidet, welche/r Experte(n) für die Bearbeitung einer Anfrage am besten geeignet sind.

DAKODA

Prompt for mixtral

```
You are an editor for speakers of German as a foreign language.
Fix the grammatical errors of the following German text.
- Proceed line by line.
- Process all lines.
- Respond only with output in German.
- Do not give any comments.
- Do not repeat the input.
- Do not use highlights / markup in your output .
```

DAKODA

Potentielle Stärke von LLMs: Kontext

- Können ganze Texte verarbeiten statt nur isolierte Sätze

Merlin 1071_0248325

- Learner input:
  Hallo Ingo, ich setze zwei zimmer eine Wohnung.
  Ich brauche nächster monat eine große der Wohnung.
  Ich möchte drei Zimmer, Küche, Bad und ein Balkon eine Wohnung.
  Bitte können Sie mir bitte helfen eine wohnungssuche?
  Bitte antworten sreiben Sie?
  Viele Grüße Eva

- Mixtral:
  Hallo Ingo, ich suche eine Wohnung mit zwei Zimmern.
  Ich brauche nexten Monat eine große Wohnung.
  Ich möchte eine Wohnung mit drei Zimmern, Küche, Bad und Balkon.
  Können Sie mir bitte bei der Wohnungssuche helfen?
  Bitte schreiben Sie antworten Sie.
  Viele Grüße Eva

- Backtranslation via pivot:
  Hallo Ingo, ich habe zwei Zimmer in eine Wohnung gestellt.
  Ich brauche im nächsten Monat eine große Wohnung.
  Ich möchte drei Zimmer, Küche, Bad und einen Balkon.
  Können Sie mir bitte helfen, eine Wohnung zu finden?
  Bitte antworten Sie?
  Viele Grüße Eva

DAKODA

Probleme bei LLMs im Chat-Modus

- Keine vollständige Verarbeitung
- Ignorieren der Anweisungen (z.B. hinsichtlich Weglassen von Kommentaren)
- Vergessen der Persona

=> roher LLM-output muss bereinigt werden

DAKODA

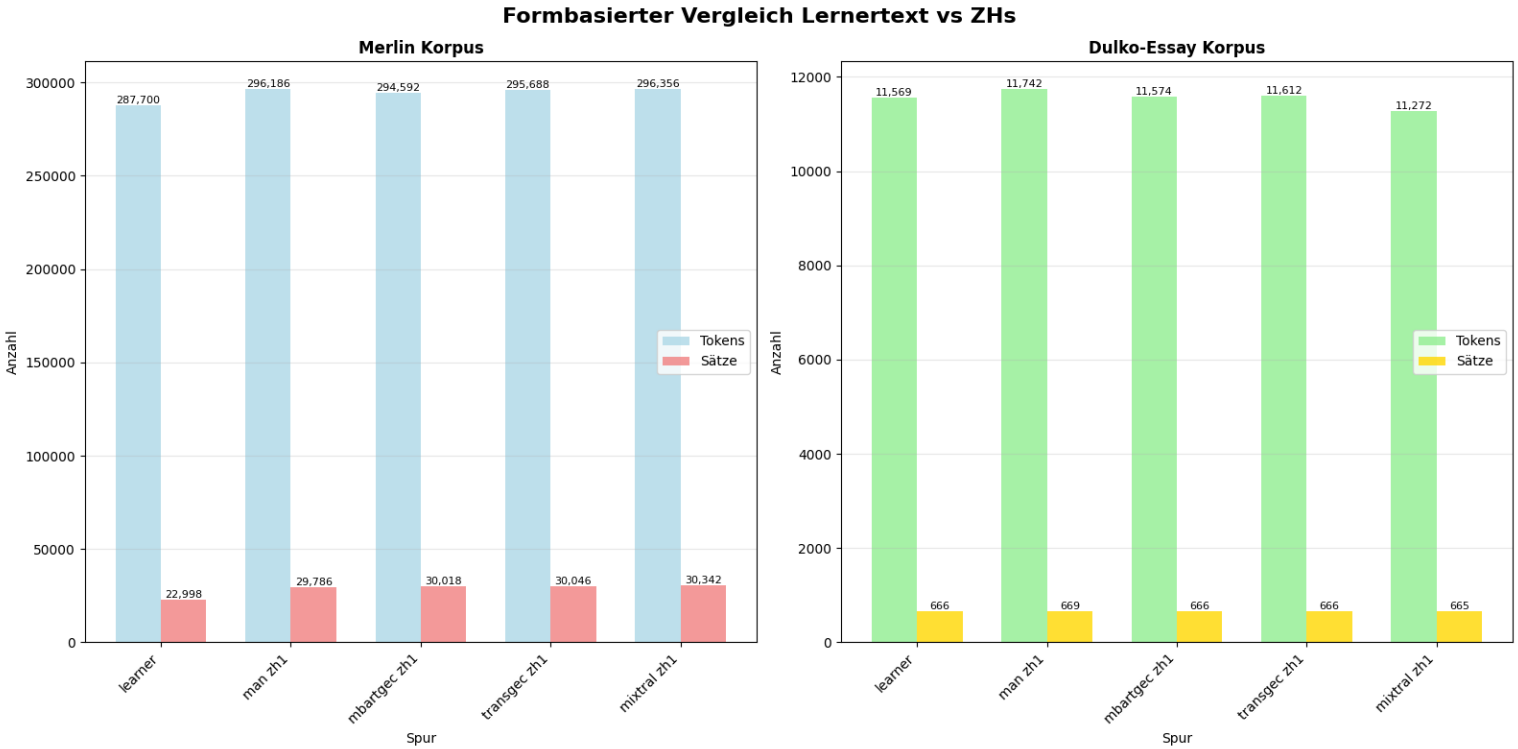
LLM Beispiel: Vergessen der Persona (Merlin 1023_0001420)

Learner text:
... Sehr geehrte Damen und Herren, ich habe Ihre Anzeige gelesen und bin sehr interesiert daran teil zu nehmen. Deshalb habe ich mich entschieden diesen Brief zu schreiben und ich hoffe sie antworten mir auf meine Fragen. Ich komme aus Kroatien und bin 28 Jahre alt. Seit ein paar Jahren lerne ich Deutsch und dass Au-pair Stelle in Deutschland wäre für mich noch eine tolle möglichkeit meine sprache zu verbessern. ...

Output von LEO:
Sehr geehrte Frau/Herr Meier, Vielen Dank für **Ihr** Interesse an **unserem** Au-pair-Programm. **Wir** freuen **uns**, dass **Sie unsere** Anzeige gefunden haben und sich dafür interessieren, an **unserem** Programm teilzunehmen. ... Mit freundlichen Grüßen, Ihr Deutsch als Fremdsprache-Lehrer

DAKODA

Formbasierter Vergleich Lernertext vs ZHs



- Merlin: ZH-Spuren haben 2-3 % mehr Tokens ; ca. 30% mehr ‘Sätze’. (Letzteres größtenteils ein Artefakt davon, wie z.B. Adress- und Grußzeilen segmentiert wurden)
- Dulko-Essays (Beeh et al. 2021): nahezu keine Unterschiede

Evaluation

Inspiration Maschinelle Übersetzung

- Maschinelle Übersetzung: automatische Übersetzung vs manuelle Referenzübersetzung(en)
- Allgemein: Ähnlichkeit einer Textversion (hypothesis, candidate) zu einer anderen (reference)
- Wir können dieselben Metriken auch auf den Fall von Lernertexten und manuellen THs anwenden.

	TH1	TH2
Learner	61.17 / 0.93	53.70 / 0.91
TH1	-/-	81.77 / 0.97

Table 2: Textähnlichkeit zwischen manuellen Textspuren in Merlin (reference layers: Spalten; bleu [Papineni et al., 2002] and bert scores [Zhang et al., 2020])

DAKODA

bleu- und bert-scores

Korpus	mbart trans	llama 3	hes-sianai leo 13b	mixtral 47b	mbart-gec	trans-gec
Merlin	60.44 / 0.92	66.23 / 0.94	56.18 / 0.91	62.81 / 0.93	93.36 / 0.98	81.64 /0.97
Dulko Essays	68.31 / 0.94	70.13 / 0.95	68.25 / 0.94	66.79 / 0.94	77.92 / 0.96	78.62 / 0.96

- Textähnlichkeit zwischen automatischen ZHs und manuell erstellten ZH1 (bleu , bert scores)
- Für mbart trans und die LLMs stellen die Werte den Durchschnitt von drei Durchläufen dar.

DAKODA

Evaluation

Überlappung in den Dependenzrelationen zwischen automatischen und manuellen THs

- Bert und Bleu-scores sind sehr oberflächenorientiert
- Wir extrahieren die Dependenzrelationen für Paare von alignierten Sätzen und messen ihre Überlappung mit Jaccard Similarity für multisets `simphile`)[1].
- Durchschnitte auf alignierten Sätzen für die beiden besten Systeme:
  - mbartgec: 0.953
  - transgec: 0.925
- Bei den bert , bleu scores lag mbartgec weiter vor transgec auf den Merlin-Daten!
- Jaccard-Wert legt nahe, dass dies wohl eher an fluency liegt als an größerer syntaktischer Ähnlichkeit

[1] https://pypi.org/project/simphile/

DAKODA

Evaluation

Task-orientiert: Verteilung von (automat.) Stufenannotationen auf den Spuren

Merlin	ADV	INV	SEP	SVO	VEND	Total
learner	0	4538 (24.7)	4333 (23.6)	6369 (34.7)	3094 (16.9)	18334
man zh1	0	4760 (25.2)	4354 (23.0)	6356 (33.6)	3428 (18.1)	18898
mbartgec zh1	0	4718 (25.1)	4360 (23.2)	6288 (33.5)	3400 (18.1)	18766
transgec zh1	0	4786 (25.4)	4375 (23.2)	6251 (33.1)	3449 (18.3)	18861
mixtral zh1	0	4803 (25.6)	4436 (23.7)	6146 (32.8)	3341 (17.8)	18726

- Insgesamt sehr große Ähnlichkeit in der Verteilung von automatischen Stufenannotation auf der Lernerspur und den ZHs.
- Aber:
  - Die Stufenannotationen in den Tabellen sind nicht unbedingt korrekt.
  - Die Stufenannotationen auf unterschiedlichen Layers betreffen nicht unbedingt dieselben (korrespondierenden) Verbinstanzen.

DAKODA

Evaluation

Task-orientiert: Stufenannotation auf alignierten Verbtokens (transgec)

- Wir betrachten die Stufenvorhersagen auf den alignierten finiten Verben zwischen transgec und Lernerspur.
- Auch hier sehr große Ähnlichkeit: Verbstellung unterscheidet sich anscheinend nicht oft zwischen Lernerspur und ZH.

Exact Match Accuracy: 0.879

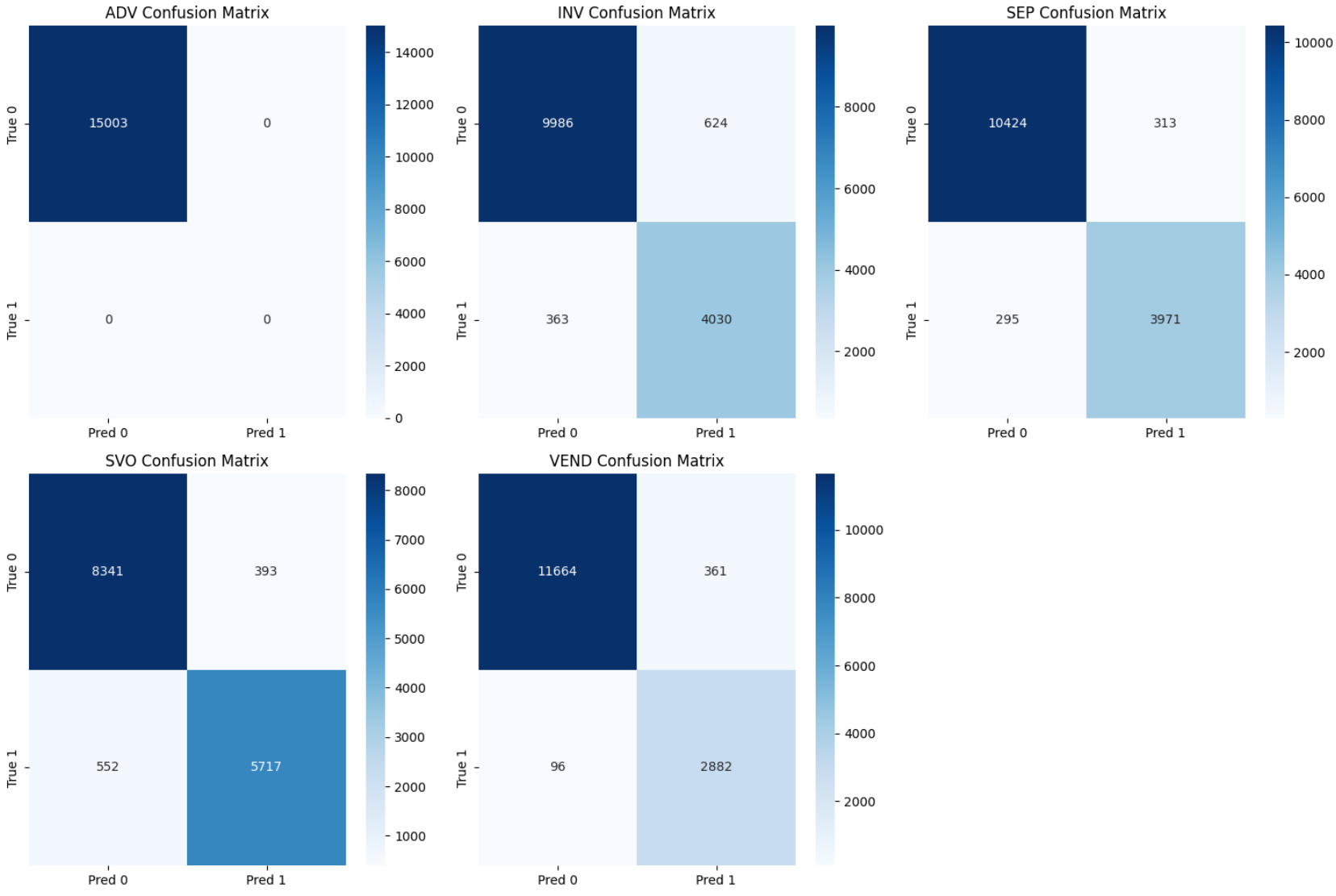
	ADV	INV	SEP	SVO	VEND
Precision	0	0.866	0.927	0.936	0.889
Recall	0	0.917	0.931	0.912	0.968
F1-Score	0	0.891	0.929	0.924	0.927
Support	0	4393	4266	6269	2978

	Micro	Macro
Precision	0.908	0.723
Recall	0.927	0.746
F1-score	0.917	0.734

DAKODA

Evaluation

Task-orientiert: Stufenannotation auf alignierten Verbtokens (transgec)



Evaluation

Task-orientiert: Stufenannotation auf alignierten Verbtokens (manuelle ZH1)

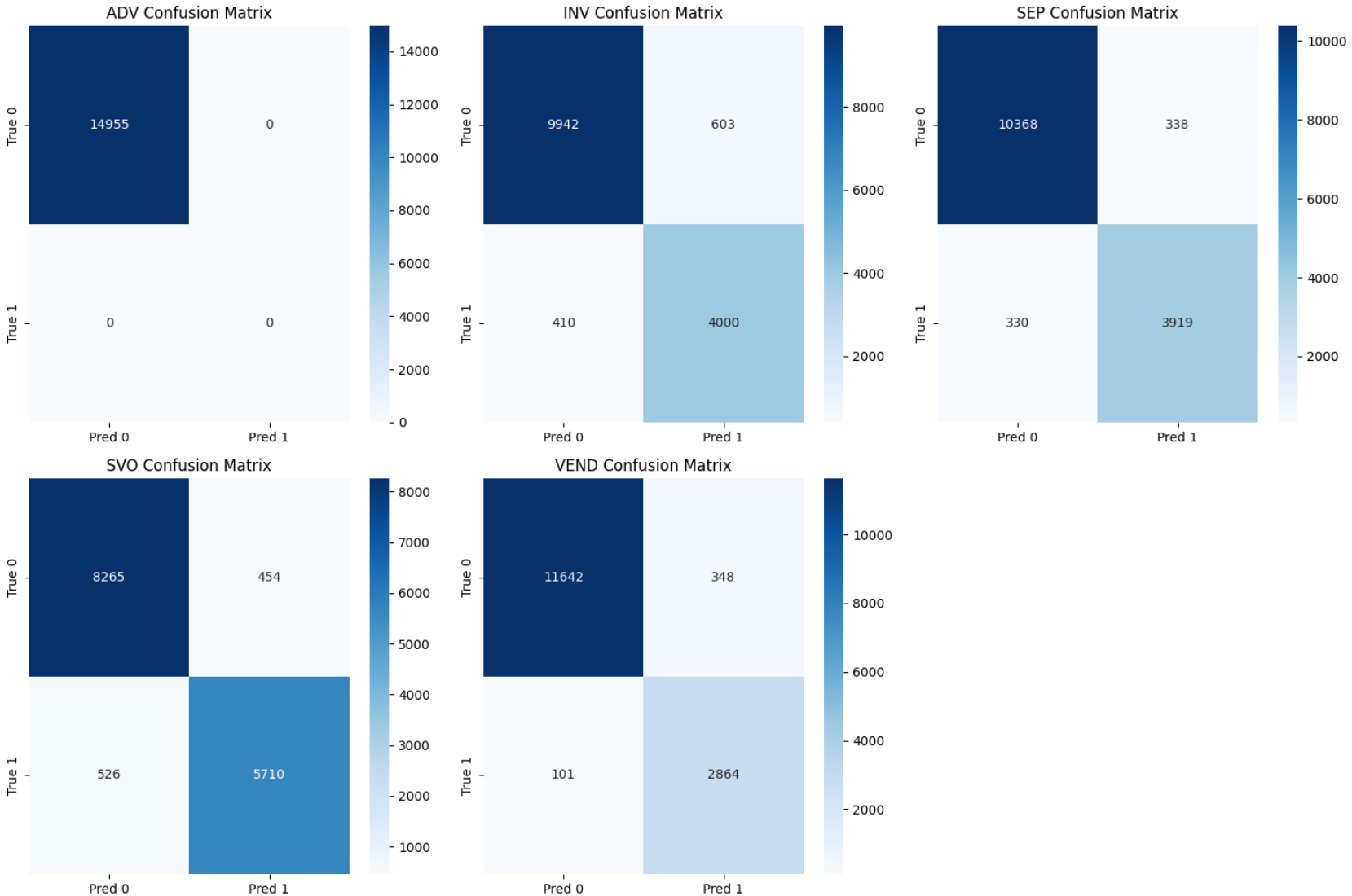
Exact Match Accuracy: 0.873

	ADV	INV	SEP	SVO	VEND
Precision	0	0.869	0.921	0.926	0.892
Recall	0	0.907	0.922	0.916	0.966
F1-Score	0	0.888	0.921	0.921	0.927
Support	0	4410	4249	6236	2965

	Micro	Macro
Precision	0.904	0.722
Recall	0.923	0.742
F1-score	0.914	0.731

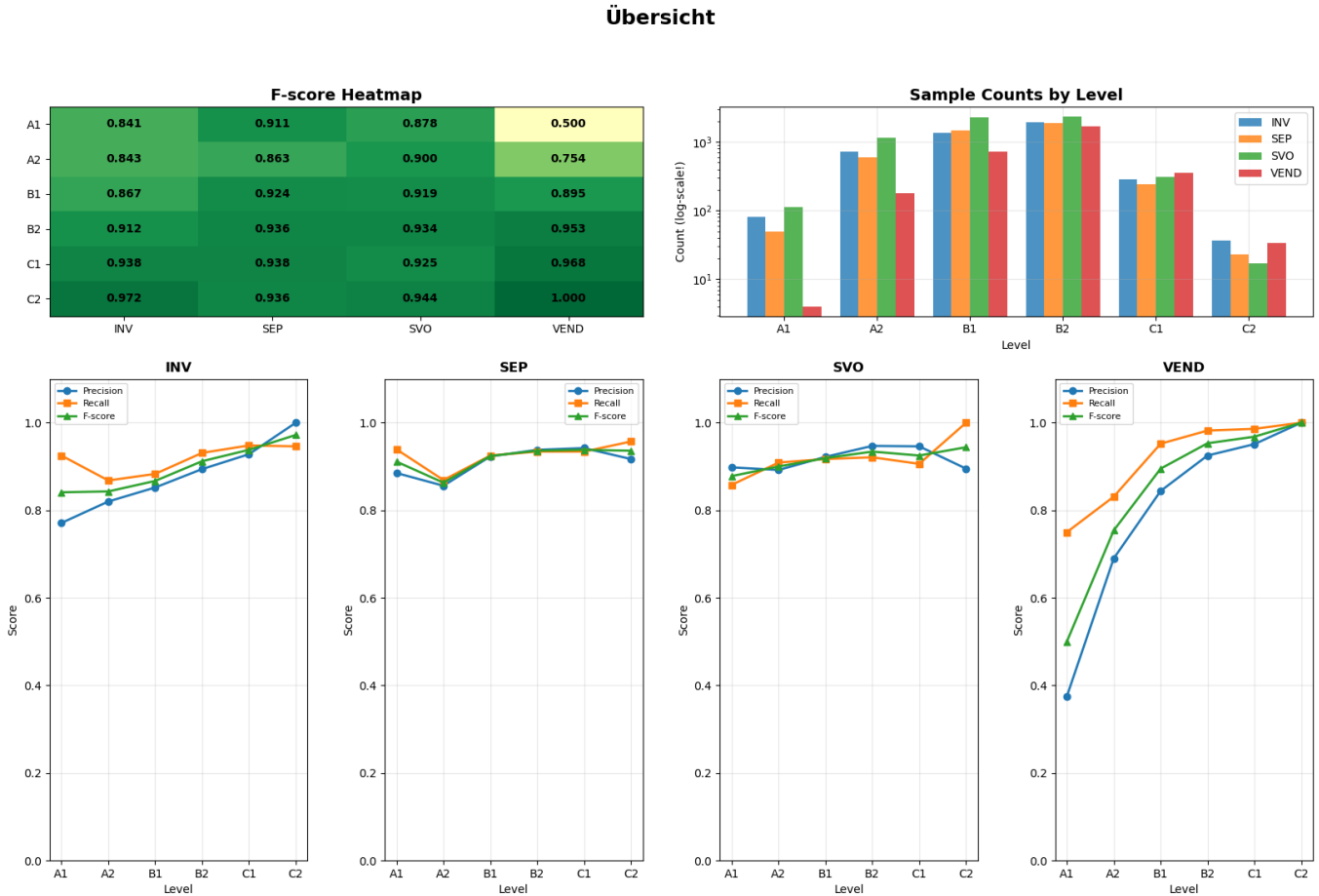
Evaluation

Task-orientiert: Stufenannotation auf alignierten Verbtokens (manuelle ZH1)



Evaluation

Task-orientiert: Stufenannotation auf alignierten Verbtokens (manuelle ZH1) pro GER-Niveau



Fazit

- Zumindest für bestimmte Textsorten möglich, automatische ZHs zu erzeugen, die manuellen ZH1 sehr ähnlich sind
- Wissenstransfer von ZH- zur Lernerparses noch nicht erfolgreich
- Hypothese: liegt an konservativem Ansatz der ‘Hybridisierung’
- Einladung, die automatischen ZHs für andere Zwecke zu nutzen und mit manuellen ZHs zu vergleichen

DAKODA

Giuseppe Attardi and Felice Dell'Orletta. Reverse revision and linear tree combination for dependency parsing. In Mari Ostendorf, Michael Collins, Shri Narayanan, Douglas W. Oard, and Lucy Vanderwende, editors, *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 261–264, Boulder, Colorado, June 2009. Association for Computational Linguistics. URL https://aclanthology.org/N09-2066/.

DAKODA

James Barry, Joachim Wagner, and Jennifer Foster. The ADAPT enhanced dependency parser at the IWPT 2020 shared task. In Gosse Bouma, Yuji Matsumoto, Stephan Oepen, Kenji Sagae, Djamé Seddah, Weiwei Sun, Anders Søgaard, Reut Tsarfaty, and Dan Zeman, editors, *Proceedings of the 16th International Conference on Parsing Technologies and the IWPT 2020 Shared Task on Parsing into Enhanced Universal Dependencies*, pages 227–235, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.iwpt-1.24. URL https://aclanthology.org/2020.iwpt-1.24/.

Christoph Beeh, Ewa Drewnowska-Vargáné, Péter Kappel, Bernadett Modrián-Horváth, Andreas Nolda, Orsolya Rauzs, and György Scheibl. *Dulko-Handbuch. Aufbau und Annotationsverfahren des deutsch-ungarischen Lernerkorpus*. Institut für Germanistik Universität Szeged, 2021. ISBN 9789633067673.

DAKODA

Robert Bley-Vroman. The comparative fallacy in interlanguage studies: The case of systematicity. *Language Learning*, 33(1): 1–17, 1983. doi: https://doi.org/10.1111/j.1467-1770.1983.tb00983.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-1770.1983.tb00983.x.

Adriane Boyd. Using Wikipedia edits in low resource grammatical error correction. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 79–84, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6111. URL https://aclanthology.org/W18-6111.

DAKODA

Adriane Boyd, Jirka Hana, Lionel Nicolas, Detmar Meurers, Katrin Wisniewski, Andrea Abel, Karin Schöne, Barbora Štindlová, and Chiara Vettori. The MERLIN corpus: Learner language and the CEFR. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1281–1288, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/606_Paper.pdf.

DAKODA

Tao Fang, Xuebo Liu, Derek F. Wong, Runzhe Zhan, Liang Ding, Lidia S. Chao, Dacheng Tao, and Min Zhang. TransGEC: Improving grammatical error correction with translationese. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3614–3633, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.223. URL https://aclanthology.org/2023.findings-acl.223.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.147. URL https://aclanthology.org/2020.findings-emnlp.147/.

Christine Köhn and Arne Köhn. An annotated corpus of picture stories retold by language learners. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 121–132. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/W18-4914.

Cédric Krummes and Astrid Ensslin. What's hard in german? whig: a british learner corpus of german. *Corpora*, 9(2):191–205, 2014. doi: 10.3366/cor.2014.0057. URL https://doi.org/10.3366/cor.2014.0057.

DAKODA

Ronja Laarmann-Quante, Katrin Ortmann, Anna Ehlert, Maurice Vogel, and Stefanie Dipper. Annotating orthographic target hypotheses in a German L1 learner corpus. In Joel Tetreault, Jill Burstein, Claudia Leacock, and Helen Yannakoudakis, editors, *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 444–456, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5051. URL https://aclanthology.org/W17-5051.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8: 726–742", 2020. doi: 10.1162/tacl_a_00343. URL https://aclanthology.org/2020.tacl-1.47/.

DAKODA

Anke Lüdeling. Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora [Ambigutities and Categorization: Problems in the Annotation of Learner Corpora]. In Maik Walter and Patrick Grommes, editors, *Fortgeschrittene Lernervarietäten*, pages 119–140. Max Niemeyer Verlag, 2008. ISBN 9783484970342. doi: doi:10.1515/9783484970342.2.119. URL https://doi.org/10.1515/9783484970342.2.119.

Stefan Müller. Mehrfache Vorfeldbesetzung. *Deutsche Sprache*, 31 (1):29–62, 2003. https://hpsg.hu-berlin.de/~stefan/Pub/mehr-vf-ds.html.

DAKODA

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040.

Ines Rehbein, Hagen Hirschmann, Anke Lüdeling, and Marc Reznicek. Better tags give better trees – or do they? *Linguistic Issues in Language Technology*, 7, Jan. 2012. doi: 10.33011/lilt.v7i.1279. URL https://journals.colorado.edu/index.php/lilt/article/view/1279.

DAKODA

Marc Reznicek, Anke Lüdeling, Cedric Krummes, Franziska Schwantuschke, Maik Walter, Karin Schmidt, Hagen Hirschmann, and Torsten Andreas. Das Falko-Handbuch. Korpusaufbau und Annotationen. Technischer bericht, Humboldt-Universität zu Berlin, 2012. URL https://hu.berlin/falkohandbuch. Version 2.01.

Marc Reznicek, Anke Lüdeling, and Hagen Hirschmann. Competing target hypotheses in the Falko corpus. In Ana Díaz-Negrillo, Paul Thompson, and Nicolas Ballier, editors, *Automatic treatment and analysis of learner corpus data*, pages 101–123. John Benjamins Publishing Company, 2013.

DAKODA

Josef Ruppenhofer, Annette Portmann, Matthias Schwendemann, Christine Renker, Katrin Wisniewski, and Torsten Zesch. Where it's at: Annotating verb placement types in learner language. In Siyao Peng and Ines Rehbein, editors, *Proceedings of the 19th Linguistic Annotation Workshop (LAW-XIX-2025)*, pages 187–200, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-262-6. doi: 10.18653/v1/2025.law-1.15. URL https://aclanthology.org/2025.law-1.15/.

Steinthor Steingrimsson, Hrafn Loftsson, and Andy Way. SentAlign: Accurate and scalable sentence alignment. In Yansong Feng and Els Lefever, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 256–263, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-demo.22. URL https://aclanthology.org/2023.emnlp-demo.22/.

DAKODA

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. URL https://arxiv.org/abs/2302.13971.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

DAKODA

Manuelle Regeln auf hybridisierten Lernerparses

Kolipsi-1 L2 (L + mbartgec ZH)

Label	precision	recall	f1-score	support
ADV	0.00	0.00	0.00	19
INV	0.91	0.94	0.92	71
SEP	0.85	0.76	0.80	80
SVO	0.85	0.92	0.88	108
VEND	0.92	0.96	0.94	68
micro avg	0.88	0.84	0.86	346
macro avg	0.70	0.72	0.71	346
weighted avg	0.83	0.84	0.83	346

DAKODA

Manuelle Regeln auf rohen Lernerparses

Kolipsi-1 L2

Label	precision	recall	f1-score	support
ADV	0.00	0.00	0.00	19
INV	0.91	0.96	0.93	71
SEP	0.82	0.79	0.80	80
SVO	0.85	0.93	0.89	108
VEND	0.92	0.96	0.94	68
micro avg	0.87	0.86	0.86	346
macro avg	0.70	0.73	0.71	346
weighted avg	0.82	0.86	0.84	346

DAKODA

Evaluation

Task-orientiert: Stufenannotation auf alignierten Verbtokens (manuelle ZH1) pro GER-Niveau [numerisch]

	Metric	ADV	INV	SEP	SVO	VEND
A1	P	0	0.771	0.885	0.898	0.375
A1	R	0	0.925	0.939	0.858	0.750
A1	F	0	0.841	0.911	0.878	0.500
A1	#	0	80	49	113	4
A2	P	0	0.820	0.856	0.892	0.690
A2	R	0	0.868	0.869	0.909	0.831
A2	F	0	0.843	0.863	0.900	0.754
A2	#	0	728	589	1141	177
B1	P	0	0.852	0.923	0.922	0.844
B1	R	0	0.883	0.925	0.917	0.952
B1	F	0	0.867	0.924	0.919	0.895
B1	#	0	1369	1471	2298	712
B2	P	0	0.894	0.938	0.947	0.925
B2	R	0	0.931	0.934	0.921	0.982
B2	F	0	0.912	0.936	0.934	0.953
B2	#	0	1909	1873	2359	1687
C1	P	0	0.928	0.942	0.946	0.951
C1	R	0	0.948	0.934	0.906	0.986
C1	F	0	0.938	0.938	0.925	0.968
C1	#	0	287	244	308	351
C2	P	0	1.000	0.917	0.895	1.000
C2	R	0	0.946	0.957	1.000	1.000
C2	F	0	0.972	0.936	0.944	1.000
C2	#	0	37	23	17	34

DAKODA

Verwandte Tasks zu GEC/Fehlerkorrektur

- Maschinelle Übersetzung
- Normalisierung historischer Daten
- Normalisierung mündlicher Daten
  - Hamwa nich und kriegen wir oooch nich wieder rein => "Haben wir nicht und kriegen wir auch nicht wieder (he)rein"
- Normalisierung von social media Daten
- Text-to-speech
  - Pasta für 1,20 Euro => "ein(en) Euro zwanzig"
  - Notenschnitt von 1,20 => "eins Komma zwei null"
- ...

DAKODA

Weitere GEC-Ansätze

- naive Übersetzung von ‘fehlerhaftem’ L2-Deutsch in L1(-näheres) Deutsch, potentiell mit expliziter Zwischenstation in einer Scharniersprache (z.B. dem Englischen)
- Korrektur durch klassisches Sprachmodell (z.B. kenlm)
- Fehler finden, dann korrigieren

- Variationen / Hilfstechniken, z.B. Verwendung synthetischer Daten (z.B. künstlich verzerrte L1 Daten; edit history von L1-Texten)

DAKODA

WTLD-Daten aus Six corpus sample

Vorhersagen nur auf dem syntaxdot-Parse

Label	precision	recall	f1-score	support
ADV	0.00	0.00	**0.00**	14
INV	0.50	0.89	0.64	9
SEP	0.90	0.82	0.86	11
SVO	0.85	0.83	0.84	82
VEND	0.74	0.93	0.82	15
———	———	——	———	———
micro avg	0.79	0.76	0.77	131
macro avg	0.60	0.69	**0.63**	131
weighted avg	0.73	0.76	0.74	131

DAKODA

WTLD-Daten aus Six corpus sample

Vorhersagen auf der Kombination von 5 Parses

Label	precision	recall	f1-score	support
ADV	0.00	0.00	0.00	14
INV	0.30	0.67	0.41	9
SEP	0.58	0.64	0.61	11
SVO	0.88	0.77	0.82	82
VEND	0.72	0.87	0.79	15
———	———	——	———	———
micro avg	0.73	0.68	0.70	131
macro avg	0.50	0.59	**0.53**	131
weighted avg	0.70	0.68	0.68	131

DAKODA

CDPS-Daten aus Six corpus sample

Vorhersagen nur auf dem syntaxdot-Parse

Label	precision	recall	f1-score	support
ADV	0.00	0.00	**0.00**	15
INV	0.41	0.38	0.40	65
SEP	0.52	0.47	0.50	57
SVO	0.34	0.39	0.36	41
VEND	0.47	0.48	0.47	48
———	———	——	———	——
micro avg	0.44	0.40	0.42	226
macro avg	0.35	0.35	**0.35**	226
weighted avg	0.41	0.40	0.41	226

DAKODA

CDPS-Daten aus Six corpus sample

Vorhersagen auf der Kombination von 5 Parses

Label	precision	recall	f1-score	support
ADV	0.00	0.00	0.00	15
INV	0.46	0.40	0.43	65
SEP	0.53	0.49	0.51	57
SVO	0.33	0.37	0.35	41
VEND	0.46	0.48	0.47	48
micro avg	0.45	0.41	0.43	226
macro avg	0.36	0.35	**0.35**	226
weighted avg	0.42	0.41	0.41	226

DAKODA

KLP1-Daten aus Six corpus sample

Vorhersagen nur auf dem syntaxdot-Parse

Label	precision	recall	f1-score	support
ADV	0.00	0.00	0.00	19
INV	0.91	0.96	0.93	71
SEP	0.82	0.79	0.80	80
SVO	0.85	0.93	0.89	108
VEND	0.92	0.96	0.94	68
———	———	———	———	———
micro avg	0.87	0.86	0.86	346
macro avg	0.70	0.73	0.71	346
weighted avg	0.82	0.86	0.84	346

DAKODA

KLP1-Daten aus Six corpus sample

Vorhersagen auf der Kombination von 5 Parses

Label	precision	recall	f1-score	support
ADV	0.00	0.00	0.00	19
INV	0.87	0.87	0.87	71
SEP	0.83	0.79	0.81	80
SVO	0.88	0.88	0.88	108
VEND	0.92	0.96	0.94	68
micro avg	0.87	0.82	0.85	346
macro avg	0.70	0.70	0.70	346
weighted avg	0.83	0.82	0.82	346

DAKODA

Stufenzuweisung mit decision tree auf anderem Datensatz

- Wir trainieren einen decision tree (SKLEARN) , der von uns definierte Merkmale betrachtet.
- Wir trainieren auf einem der drei Datensätze und wenden den gelernten decision tree dann auf die anderen beiden Datensätze an.

DAKODA

Stufenzuweisung mit decision tree auf anderem Datensatz

Training: Six corpus sample; Test: Kalibrierungssample

	precision	recall	f1-score	support
ADV	0.06	0.06	0.06	34
INV	0.81	0.70	0.75	470
SEP	0.75	0.76	0.75	327
SVO	0.90	0.73	0.81	537
VEND	0.86	0.72	0.79	418
:————	:————	:———	:————	:———–
micro avg	0.82	0.71	0.76	1786
macro avg	0.68	0.59	**0.63**	1786
weighted avg	0.83	0.71	0.76	1786

DAKODA

Stufenzuweisung mit decision tree auf anderem Datensatz

Training: Six corpus sample, Test: UDTopoADV

Label	precision	recall	f1-score	support
ADV	0.19	0.12	0.15	65
INV	0.62	0.62	0.62	234
SEP	0.61	0.80	0.70	177
SVO	0.62	0.68	0.65	254
VEND	0.78	0.80	0.79	191
micro avg	0.63	0.67	0.65	921
macro avg	0.57	0.60	**0.58**	921
weighted avg	0.62	0.67	0.65	921

DAKODA

Stufenzuweisung mit decision tree auf anderem Datensatz

Training: UDTopoADV; Test: Six corpus sample

Label	precision	recall	f1-score	support
ADV	0.23	0.19	0.21	59
INV	0.60	0.86	0.71	297
SEP	0.88	0.80	0.84	292
SVO	0.80	0.68	0.74	422
VEND	0.75	0.83	0.79	246
———	———	———	———	———
micro avg	0.72	0.75	0.74	1316
macro avg	0.65	0.67	**0.65**	1316
weighted avg	0.74	0.75	0.74	1316

DAKODA

Stufenzuweisung mit decision tree auf anderem Datensatz

Training: UDTopoADV; Test: Kalibrierungssample

Label	precision	recall	f1-score	support
ADV	0.19	0.15	0.17	34
INV	0.73	0.90	0.81	470
SEP	0.96	0.85	0.90	327
SVO	0.88	0.73	0.80	537
VEND	0.89	0.89	0.89	418
———	—–	——	——	—–
micro avg	0.84	0.82	0.83	1786
macro avg	0.73	0.70	**0.71**	1786
weighted avg	0.85	0.82	0.83	1786

DAKODA

Stufenzuweisung mit decision tree auf anderem Datensatz

Training: Kalibrierungssample; Test: Six corpus sample

label	precision	recall	f1-score	support
ADV	0.43	0.10	0.16	59
INV	0.67	0.89	0.77	297
SEP	0.84	0.82	0.83	292
SVO	0.80	0.83	0.82	422
VEND	0.74	0.76	0.75	246
———	———	———	———	———
micro avg	0.76	0.80	0.77	1316
macro avg	0.69	0.68	**0.66**	1316
weighted avg	0.75	0.80	0.76	1316

DAKODA

Stufenzuweisung mit decision tree auf anderem Datensatz

Training: Kalibrierungssample; Test: UDTopoADV

label	precision	recall	f1-score	support
ADV	0.00	0.00	0.00	65
INV	0.66	0.90	0.76	234
SEP	0.70	0.85	0.77	177
SVO	0.69	0.88	0.77	254
VEND	0.77	0.83	0.80	191
———	———	———	———	———
micro avg	0.70	0.81	0.75	921
macro avg	0.56	0.69	**0.62**	921
weighted avg	0.65	0.81	0.72	921

DAKODA

Suche nach effektiveren Klassifikatoren

In-domain auf dem Six corpus sample

- Wir trainieren und testen innerhalb eines Datensatzes mit der Modellsuche von Autogluon.
- Train-Test split: 80-20
- Wir können die Ergebnisse unseres besten decision trees übertreffen.
- Für unterschiedliche Label sind unterschiedliche Klassifikatoren am besten.
- Alle sind Ensemble-Klassifikatoren.

Label	Prec	Rec	F1	MS F1	Klassifikator
ADV	0.250	0.455	0.323	1.000	WeightedEnsemble_L2
INV	0.771	0.673	0.718	1.000	LightGBMXT_BAG_L1
SEP	0.952	0.741	0.833	0.952	WeightedEnsemble_L3
SVO	0.866	0.690	0.768	1.000	LightGBM_BAG_L1
VEND	0.776	0.760	0.768	1.000	CatBoost_BAG_L1

Table 5: Linke Hälfte: decision trees im besten ermittelten Setting; rechte Spalten: Ergebnisse der Modellsuche mit Autogluon

DAKODA

Suche nach effektiveren Klassifikatoren

In-domain auf Kalibrierungssample

- Wir können die hohen Ergebnisse unseres besten decision trees übertreffen.
- ADV erreicht aber dennoch keine guten Werte.

Label	Prec	Rec	F1	MS F1	Klassifikator
ADV	0.500	0.154	0.235	0.471	WeightedEnsemble_L3
INV	0.936	0.978	0.957	0.989	WeightedEnsemble_L2
SEP	0.935	0.892	0.913	0.977	WeightedEnsemble_L3
SVO	0.935	0.893	0.913	0.983	LightGBMXT_BAG_L1
VEND	0.964	0.976	0.970	1.000	CatBoost_BAG_L1

DAKODA

Suche nach effektiveren Klassifikatoren

In domain auf UDTopoADV

- Wir können die hohen Ergebnisse des besten decision trees übertreffen.
- Alle gewählten Klassifikatoren sind Ensemble-Klassifikatoren.
- Offen ist, wie gut die Performanz der Klassifikatoren bei Anwendung auf andere Datensätze ist.

Label	Prec	Rec	F1	MS F1	Klassifikator
ADV	1.000	0.889	0.941	1.000	WeightedEnsemble_L2
INV	0.929	0.945	0.937	0.982	RandomForestEntr_BAG_L1
SEP	0.971	0.971	0.971	1.000	RandomForestEntr_BAG_L1
SVO	0.822	0.841	0.831	0.932	NeuralNetFastAI_BAG_L1
VEND	1.000	1.000	1.000	1.000	LightGBM_BAG_L1

Table 6: Linke Hälfte: decision trees im besten ermittelten Setting; rechte Spalten: Ergebnisse der Modellsuche mit Autogluon

DAKODA

Suche nach bestem Parser

Task-orientierte Evaluation:syntaxdot

- Ergebnisse der Stufenvorhersage mit händischen Regeln auf den syntaxdot-Parses für die Lernerspur von **CDPS** (Six Corpus sample)

Label	precision	recall	f1-score	support
ADV	0.00	0.00	0.00	15
INV	0.41	0.38	0.40	65
SEP	0.52	0.47	0.50	57
SVO	0.34	0.39	0.36	41
VEND	0.47	0.48	0.47	48
micro avg	0.44	0.40	0.42	226
macro avg	0.35	0.35	0.35	226
weighted avg	0.41	0.40	0.41	226

DAKODA

Suche nach bestem Parser

Task-orientierte Evaluation:stanza

- Ergebnisse der Stufenvorhersage mit händischen Regeln auf den stanza-Parses für die Lernerspur von **CDPS** (Six Corpus sample)

	precision	recall	f1-score	support
ADV	0.00	0.00	0.00	15
INV	0.46	0.40	0.43	65
SEP	0.55	0.49	0.52	57
SVO	0.32	0.34	0.33	41
VEND	0.42	0.44	0.43	48
———	———	———	———	———
micro avg	0.44	0.39	0.42	226
macro avg	0.35	0.33	0.34	226
weighted avg	0.42	0.39	0.40	226

DAKODA

Suche nach bestem Parser

Task-orientierte Evaluation: trankit

- Ergebnisse der Stufenvorhersage mit händischen Regeln auf den trankit-Parses für die Lernerspur von **CDPS** (Six corpus sample)

	precision	recall	f1-score	support
ADV	0.00	0.00	0.00	15
INV	0.44	0.40	0.42	65
SEP	0.53	0.49	0.51	57
SVO	0.35	0.39	0.37	41
VEND	0.46	0.48	0.47	48
micro avg	0.45	0.41	0.43	226
macro avg	0.36	0.35	0.35	226
weighted avg	0.42	0.41	0.42	226

DAKODA

Suche nach bestem Parser

Task-orientierte Evaluation: udpipe1

- Ergebnisse der Stufenvorhersage mit händischen Regeln auf den udpipe1-Parses für die Lernerspur von **CDPS** (Six corpus sample)

	precision	recall	f1-score	support
ADV	0.00	0.00	0.00	15
INV	0.47	0.42	0.44	65
SEP	0.52	0.49	0.50	57
SVO	0.36	0.34	0.35	41
VEND	0.40	0.35	0.37	48
micro avg	0.45	0.38	0.41	226
macro avg	0.35	0.32	0.33	226
weighted avg	0.42	0.38	0.40	226

DAKODA

Suche nach bestem Parser

Task-orientierte Evaluation: udpipe2

- Ergebnisse der Stufenvorhersage mit händischen Regeln auf den udpipe2-Parses für die Lernerspur von **CDPS** (Six corpus sample)

	precision	recall	f1-score	support
ADV	0.00	0.00	0.00	15
INV	0.44	0.42	0.43	65
SEP	0.53	0.47	0.50	57
SVO	0.33	0.34	0.34	41
VEND	0.46	0.48	0.47	48
micro avg	0.45	0.40	0.42	226
macro avg	0.35	0.34	0.35	226
weighted avg	0.42	0.40	0.41	226

DAKODA

Suche nach bestem Parser

Task-orientierte Evaluation: syntaxdot

- Ergebnisse der Stufenvorhersage mit händischen Regeln auf den syntaxdot-Parses für die Lernerspur von **KLP1** (Six corpus sample)

	precision	recall	f1-score	support
ADV	0.00	0.00	0.00	19
INV	0.91	0.96	0.93	71
SEP	0.82	0.79	0.80	80
SVO	0.85	0.93	0.89	108
VEND	0.92	0.96	0.94	68
micro avg	0.87	0.86	0.86	346
macro avg	0.70	0.73	0.71	346
weighted avg	0.82	0.86	0.84	346

DAKODA

Suche nach bestem Parser

Task-orientierte Evaluation: stanza

- Ergebnisse der Stufenvorhersage mit händischen Regeln auf den stanza-Parses für die Lernerspur von **KLP1** (Six corpus sample)

	precision	recall	f1-score	support
ADV	0.00	0.00	0.00	19
INV	0.83	0.85	0.84	71
SEP	0.82	0.69	0.75	80
SVO	0.87	0.84	0.85	108
VEND	0.89	0.94	0.91	68
micro avg	0.85	0.78	0.82	346
macro avg	0.68	0.66	0.67	346
weighted avg	0.81	0.78	0.79	346

DAKODA

Suche nach bestem Parser

Task-orientierte Evaluation: trankit

- Ergebnisse der Stufenvorhersage mit händischen Regeln auf den trankit-Parses für die Lernerspur von **KLP1** (Six corpus sample)

	precision	recall	f1-score	support
ADV	0.00	0.00	0.00	19
INV	0.86	0.85	0.85	71
SEP	0.85	0.79	0.82	80
SVO	0.88	0.86	0.87	108
VEND	0.93	0.96	0.94	68
micro avg	0.88	0.81	0.84	346
macro avg	0.70	0.69	0.70	346
weighted avg	0.83	0.81	0.82	346

DAKODA

Suche nach bestem Parser

Task-orientierte Evaluation: udpipe1

- Ergebnisse der Stufenvorhersage mit händischen Regeln auf den udpipe1-Parses für die Lernerspur von **KLP1** (Six corpus sample)

	precision	recall	f1-score	support
ADV	0.00	0.00	0.00	19
INV	0.71	0.77	0.74	71
SEP	0.79	0.75	0.77	80
SVO	0.89	0.77	0.83	108
VEND	0.78	0.94	0.85	68
micro avg	0.80	0.76	0.78	346
macro avg	0.64	0.65	0.64	346
weighted avg	0.76	0.76	0.76	346

DAKODA

Suche nach bestem Parser

Task-orientierte Evaluation: udpipe2

- Ergebnisse der Stufenvorhersage mit händischen Regeln auf den udpipe2-Parses für die Lernerspur von **KLP1** (Six corpus sample)

	precision	recall	f1-score	support
ADV	0.00	0.00	0.00	19
INV	0.83	0.85	0.84	71
SEP	0.82	0.69	0.75	80
SVO	0.87	0.84	0.85	108
VEND	0.89	0.94	0.91	68
micro avg	0.85	0.78	0.82	346
macro avg	0.68	0.66	0.67	346
weighted avg	0.81	0.78	0.79	346

DAKODA

Suche nach bestem Parser

Task-orientierte Evaluation: syntaxdot

- Ergebnisse der Stufenvorhersage mit händischen Regeln auf den syntaxdot-Parses für die Lernerspur von **SWIKO-WTLD** (Six corpus sample)

	precision	recall	f1-score	support
ADV	0.00	0.00	0.00	14
INV	0.50	0.89	0.64	9
SEP	0.90	0.82	0.86	11
SVO	0.85	0.83	0.84	82
VEND	0.74	0.93	0.82	15
micro avg	0.79	0.76	0.77	131
macro avg	0.60	0.69	0.63	131
weighted avg	0.73	0.76	0.74	131

DAKODA

Suche nach bestem Parser

Task-orientierte Evaluation: stanza

- Ergebnisse der Stufenvorhersage mit händischen Regeln auf den stanza-Parses für die Lernerspur von **SWIKO-WTLD** (Six corpus sample)

	precision	recall	f1-score	support
ADV	0.00	0.00	0.00	14
INV	0.30	0.78	0.44	9
SEP	0.47	0.64	0.54	11
SVO	0.89	0.76	0.82	82
VEND	0.52	0.80	0.63	15
micro avg	0.67	0.67	0.67	131
macro avg	0.44	0.59	0.48	131
weighted avg	0.67	0.67	0.66	131

DAKODA

Suche nach bestem Parser

Task-orientierte Evaluation: trankit

- Ergebnisse der Stufenvorhersage mit händischen Regeln auf den trankit-Parses für die Lernerspur von **SWIKO-WTLD** (Six corpus sample)

	precision	recall	f1-score	support
ADV 0.00	0.00	0.00	14	
INV	0.30	0.67	0.41	9
SEP	1.00	0.64	0.78	11
SVO	0.85	0.71	0.77	82
VEND	0.72	0.87	0.79	15
micro avg	0.74	0.64	0.69	131
macro avg	0.58	0.58	0.55	131
weighted avg	0.72	0.64	0.67	131

DAKODA

Suche nach bestem Parser

Task-orientierte Evaluation: udpipe1

- Ergebnisse der Stufenvorhersage mit händischen Regeln auf den udpipe1-Parses für die Lernerspur von **SWIKO-WTLD** (Six corpus sample)

precision	recall	f1-score	support	
ADV	0.00	0.00	0.00	14
INV	0.30	0.78	0.44	9
SEP	0.36	0.45	0.40	11
SVO	0.90	0.65	0.75	82
VEND	0.38	0.73	0.50	15
micro avg	0.61	0.58	0.59	131
macro avg	0.39	0.52	0.42	131
weighted avg	0.66	0.58	0.59	131

DAKODA

Suche nach bestem Parser

Task-orientierte Evaluation: udpipe2

- Ergebnisse der Stufenvorhersage mit händischen Regeln auf den udpipe2-Parses für die Lernerspur von **SWIKO-WTLD** (Six corpus sample)

	precision	recall	f1-score	support
ADV	0.00	0.00	0.00	14
INV	0.30	0.78	0.44	9
SEP	0.47	0.64	0.54	11
SVO	0.89	0.76	0.82	82
VEND	0.52	0.80	0.63	15
micro avg	0.67	0.67	0.67	131
macro avg	0.44	0.59	0.48	131
weighted avg	0.67	0.67	0.66	131

DAKODA

Majority vote

Auf dem six corpus sample

- Voraussagen von 5 Parsern
- 5-fold cross-validation auf dem 6-corpus sample
- weighted voting (gelernt nur auf den vorhergesagten Labeln)

```
ADV: F1=0.000±0.000, Acc=0.952±0.031
SVO: F1=0.827±0.031, Acc=0.873±0.019
INV: F1=0.805±0.023, Acc=0.896±0.007
SEP: F1=0.802±0.050, Acc=0.900±0.022
VEND: F1=0.858±0.047, Acc=0.937±0.020
F1 Macro: 0.658 ± 0.023
```

DAKODA