

Organisatorisches zum Softwareprojekt

Jakob Schuster

Computerlinguistik
Universität Heidelberg

Sommersemester 2026



UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386



Casey Wright
@WrightToLife



Welcome to college. Where every single person is smarter than you except for the 3 people in your group project.

Überblick

Herzlich Willkommen zur Begleitveranstaltung zum Softwareprojekt

- Teilnahmevoraussetzung:
 - Erfolgreiche Teilnahme an ECL, Prog. II, Statistik
 - Bestehen der Orientierungsprüfung
- Dauer: 1 Semester
- 6 LP + 4 LP ÜK → ca. "20h/Woche"

Heute

1. Inhalte & Ziele
2. Semesterplan
3. Benotung
4. Themen

Im Softwareprojekt soll eine computerlinguistische Aufgabenstellung

- ... weitgehend eigenverantwortlich und in Teamarbeit
- ... geplant,
- ... programmiert,
- ... getestet,
- ... dokumentiert und
- ... präsentiert werden.

Am Ende des Softwareprojekts können Sie...

1. aus einem theoretischen Ansatz einen konkreten Projektplan erarbeiten;
2. den Plan in Teamarbeit umsetzen;
3. Ihre Projektergebnisse präsentieren und analysieren.

Am Ende des Softwareprojekts können Sie...

1. aus einem theoretischen Ansatz einen konkreten Projektplan erarbeiten;
 - den Ansatz in eigenen Worten wiedergeben;
 - zu implementierende Module festlegen und Abhängigkeiten zwischen den Modulen definieren;
 - aus dem Modulplan einen Zeitplan und eine Aufgabenverteilung entwickeln;
 - Programmarchitektur, Datenstrukturen und Schnittstellen festlegen.
2. den Plan in Teamarbeit umsetzen;
3. Ihre Projektergebnisse präsentieren und analysieren.

Am Ende des Softwareprojekts können Sie...

1. aus einem theoretischen Ansatz einen konkreten Projektplan erarbeiten;
2. den Plan in Teamarbeit umsetzen;
3. Ihre Projektergebnisse präsentieren und analysieren.

Am Ende des Softwareprojekts können Sie...

1. aus einem theoretischen Ansatz einen konkreten Projektplan erarbeiten;
2. den Plan in Teamarbeit umsetzen;
 - die Ihnen zugewiesenen Module plangemäß implementieren;
 - Module so testen und dokumentieren, dass andere Teammitglieder sie verwenden können;
 - gemeinsame Instrumente entwickeln und einsetzen, die sicherstellen, dass Sie den Plan einhalten, und dass Sie Fragen und Probleme rechtzeitig klären.
3. Ihre Projektergebnisse präsentieren und analysieren.

Am Ende des Softwareprojekts können Sie...

1. aus einem theoretischen Ansatz einen konkreten Projektplan erarbeiten;
2. den Plan in Teamarbeit umsetzen;
3. Ihre Projektergebnisse präsentieren und analysieren.

Am Ende des Softwareprojekts können Sie...

1. aus einem theoretischen Ansatz einen konkreten Projektplan erarbeiten;
2. den Plan in Teamarbeit umsetzen;
3. Ihre Projektergebnisse präsentieren und analysieren.
 - Ergebnisse und Methoden so präsentieren und demonstrieren, dass andere Teams verstehen, was Sie gemacht haben;
 - Schwächen Ihrer Implementierung und Probleme beim Projektablauf identifizieren und Verbesserungsvorschläge/ Gründe angeben.

Teamarbeit

“gemeinsame Instrumente entwickeln und einsetzen, die sicherstellen, dass Sie den Plan einhalten, und dass Sie Fragen und Probleme rechtzeitig klären.”

- regelmäßige Treffen - in der Gruppe und mit dem Betreuer
- gemeinsames “Logbuch” (Wiki oder gemeinsames Dokument)
- Kommunikation über Chat, Slack, ...
- gemeinsamer Terminkalender
- wöchentliche Statusreports aller Teilnehmer
- Codereviews
- Online oder Offline
- Einsetzen eines/r Projektmanagers/in

Semesterplan

Datum	Inhalt
14.04.	Organisation, Projektvorstellung (Plenum)
Ab 27.04	Statusmeetings
12.05.	Besprechung Forschungsplan (Einreichung: 10.05.)
...	Statusmeetings
19.05.	Spezifikationsvortrag (Plenum)
...	Statusmeetings
21.07.	Abschlussvorträge, Demos (Plenum)
23.08. (So)	Projektgabe

Einteilung

Projektanmeldung

Bis 19.04. 18 Uhr, E-Mail an schuster@c1.uni-heidelberg.de mit

1. Betreff: “[SWP] Projektanmeldung”
2. mit Prioritäten zu den Projektthemen (als Rangfolge)
3. Falls Sie im Vorfeld eine Gruppe bilden konnten, die Namen der anderen Gruppenmitglieder.
4. Programmierkenntnisse in Python: (1: sehr gut, 2: gut-mittel, 3: schwierig)
5. Kenntnisse und Erfahrung in Statistik, maschinellem Lernen, neuronalen Netzen
6. Fachrichtung und Fachsemester
7. Euer ICL Username

Einteilung

Gruppeneinteilung

- 3-6 Mitglieder je Gruppe
- Vorgabe (Vorschlag) je nach Projekt
- Gruppen können im Vorfeld gebildet werden

Forschungsplan

Forschungsplan: Abgabe 10.05.2026

- Erste, übergreifende Projektbeschreibung: Ziel, Methode, Evaluierung, Daten, Tools
- Beschreibt die Problemstellung (Ziel) und den Lösungsansatz (Methode) in eigenen Worten!
- Wie sollen die Ergebnisse evaluiert werden?
- Welche Tools und Daten sollen verwendet werden?
- Abgabe schriftlich, ca. 2 Seiten
- Besprechung 12.05.

Spezifikation

Spezifikationsvortrag: 19.05.2026

- Inhaltliche Spezifikation:
 - Problemstellung, Lösungsansatz, Evaluation (← Forschungsplan)
 - Auswahl/Sichtung benötigter Ressourcen und Verfahren/Algorithmen
- Modularisierung und Aufgabenverteilung:
 - Definition von Modulen/Aufgaben und Abhängigkeiten
 - Zuordnung von Namen zu Aufgaben
 - Zeitplan (mit Parallelisierung!)
- Konkrete Planung von Programmarchitektur, Datenstrukturen, Schnittstellen
- Vortrag im Plenum, maximal 20-30 Minuten

Statusmeetings

Statusmeetings

- Individuelle 1h lange Termine für jede Gruppe mit Betreuer.
- 2 Gruppen während der Seminarzeit, fixe Termine werden gemeinsam mit den Gruppen in der nächsten Woche festgelegt
- Immer in INF325, Raum 109
- Come prepared!

Abschlussvortrag

Abschlussvortrag: 21.07.2026

- Kurze Darstellung von Ziel und Methode
- Konkrete Umsetzung der Methode
- Präsentation der Evaluationsergebnisse
- Demo
- Lessons Learned: Identifikation von Schwächen eurer Implementierung und Problemen beim Projektablauf → Verbesserungsvorschläge/ Gründe angeben.
- Vortrag im Plenum, maximal 25 Minuten + Fragen

Leistungsnachweis

Abgabe: 23.08.2026 – keine Ausnahmen möglich!

Voraussetzung für einen Schein:

- Teilnahme an allen Veranstaltungen (Plenum und Statusmeetings) (Fehlen nur mit Attest oder Absprache mit Dozenten; max. 1x unentschuldigt)
- Forschungsplan
- Spezifikationsvortrag
- Abschlussvortrag (+ Demo, sofern sinnvoll)
- Dokumentation und Archivierung des Projekts
 - Dokumentation des Quellcodes
 - README mit Installationshinweisen
 - Schriftlicher Ergebnisbericht (je nach Projekt, ca. 5 Seiten), ggf. zusätzliche Anleitungen/Userguide
 - Lizenzierung (falls der Code veröffentlicht werden soll)

Benotung

Benotungskriterien:

1. Sorgfältige Herangehensweise bei Planung, Implementierung, Dokumentation
→ Evaluation durch Betreuer
2. Präsentation des fertigen Projekts und Lessons Learned
→ Evaluation durch Betreuer und evtl. andere Projektgruppen
3. Zusammenarbeit im Team
→ Evaluation u.a. durch Teammitglieder

In Sonderfällen können Einzelnoten vergeben werden!

Vorstellung der Themen

1. Source Credibility: Repetition, Plausibility and Mitigation
2. Data2Text Generation für Fußball-Spielberichte
3. Toy-Setting für Anchor-Token-Hypothesis von Multilingual LLMs
4. Adjective Ordering Preferences und Verletzung dieser in LLMs

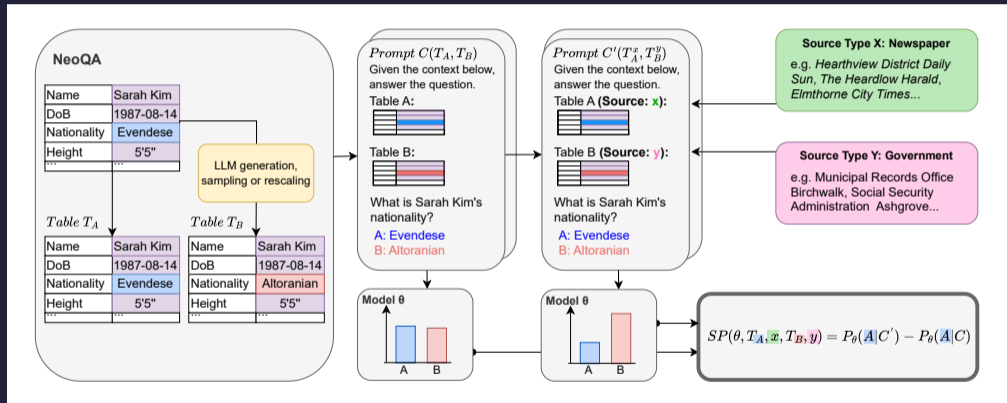
Knowledge-Conflicts: Source Credibility

(≤ 5 Personen)

- Weiterführung von Schuster, Gautam, and Markert (2026)
- LLMs haben latente Quellenpräferenzen
- Aber: Einfache Wiederholung einer nichtpräferierten Quelle überwiegt diesen Effekt
- Token-Probability basierte Metrik für gleichplausible Konflikte
- Mitigation Ansatz mit Knowledge Distillation

Knowledge-Conflicts: Source Credibility

(≤ 5 Personen)



-
- RQ1: Wie interagiert dies mit unterschiedlich plausiblen Konflikten?
- RQ2: Funktioniert dies auch bei unconstrained Generation?

Data2Text Generation für Fußball-Spielberichte

(≤ 6 Personen)

- **Data2Text:**
 - Input: Strukturierte, tabellarische Daten
 - Output: Natürlichsprachliche Verbalisierung dieser
- Umsetzung jedoch sehr individuell und an Daten und Texte angepasst
- Hier:
 - Sehr detaillierte Eventdaten über 1. Fußball Bundesliga Daten (2024)
 - Spielberichte des Kickers
- Empfohlene Literatur: Reiter (2007) und Wiseman, Shieber, and Rush (2017)

Data2Text Generation für Fußball-Spielberichte

(≤ 6 Personen)

Input (Impact Daten)

```
"index": 991, "id": 4614762957, "gameTime": "gameTime": "29:06.8300", "gameTimeInSec": 1746.83 , "squadId": 432, "player": "id": 1457, "position": "RIGHT WINGBACK DEFENDER", "positionSide": "RIGHT" , "pressure": 30.0, "actionType": "SHOT", "start": "coordinates": 31.0, 3.7, "packingZone": "DMC", "pitchPosition": "FINAL THIRD", "lane": "CENTER" , "end": "coordinates": "x": 52.5, "y": -2.4, "packingZone": null, "pitchPosition": "OPPONENT BOX", "lane": "CENTER" , "action": "MID RANGE SHOT", "phase": "SET PIECE", "shot": "distance": 21.8, "angle": 9.8, "targetPoint": null, "bodyPart": "FOOT RIGHT", "pxT": "team": 0.022565, "opponent": 0.000508, "distanceToGoal": 21.8, "result": "FAIL", "periodId": 1, "formation": "team": "4-1-3-2", "opponent": "4-3-3", "setPiece": "id": 45791855, "subPhaseId": 48959236, "mainEvent": false
```

Data2Text Generation für Fußball-Spielberichte

(≤ 6 Personen)

Output Gescrapte kicker Artikel

- Einfache Zusammenfassungen:
 - *“Can verwandelte den Elfmeter sicher zum 2:0 (15.)”*
- Figurative Sprache:
 - *“Grillitsch manövrierte eine Bogenlampe aus etwa 55 Metern ins Kölner Tor (28.)”*
 - *“Guirassy nagelte die Kugel aus 17 Metern ins rechte obere Eck (32.)”*

Toy-Setting für die Anchor-Token-Hypothese

(≤ 5 Personen)

- Mehrsprachige Sprachmodelle können Wissen über Sprachgrenzen hinweg transferieren
- **Anchor-Token-Hypothese:** Token, die in mehreren Sprachen identisch vorkommen ermöglichen diesen Transfer
- Kontrollierten Setting um sprachübergreifenden Transfer systematisch zu überprüfen

Toy-Setting für die Anchor-Token-Hypothese

(≤ 5 Personen)

- **Vorgehen:**
 - Erstellung von Korpora kleiner Kunstsprachen (mit Knowledge Base, (P)CFG o. Ä.)
 - RQ orientiertes Variieren der Korpora
 - Training und Evaluation des Transfers
- Empfohlene Literatur: Dufter and Schütze (2020)

Adjective Ordering Preferences in LLMs





(≤ 5 Personen)


- **Adjective Ordering Preference:** bevorzugte Reihenfolge von Adjektiven
 - *a big red box*
 - **a red big box*
- Aus pragmatischen Gründen (Betonung, Kontrast) kann diese Präferenz jedoch aufgeweicht werden
- RQ: Zeigen auch LLMs diese pragmatischen Ausnahmen?

Adjective Ordering Preferences in LLMs

(≤ 5 Personen)

- **Vorgehen:**
 - Entwicklung einer Benchmark zur systematischen Evaluation
 - Explorative Erweiterungen (VLLMs, Sprachen, andere Präferenzen)
- Empfohlene Literatur: Leivada 2022

-  Dufter, Philipp and Hinrich Schütze (2020). “Identifying elements essential for BERT’s multilinguality”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4423–4437.
-  Leivada, Evelina (2022). “Determining the cognitive biases behind a viral linguistic universal: the order of multiple adjectives”. In: *Humanities and Social Sciences Communications* 9.1, p. 436.
-  Reiter, Ehud (2007). “An architecture for data-to-text systems”. In: *proceedings of the eleventh European workshop on natural language generation (ENLG 07)*, pp. 97–104.
-  Schuster, Jakob, Vagrant Gautam, and Katja Markert (2026). *Whose Facts Win? LLM Source Preferences under Knowledge Conflicts*. arXiv: 2601.03746 [cs.CL]. URL: <https://arxiv.org/abs/2601.03746>.

-  Wiseman, Sam, Stuart Shieber, and Alexander Rush (Sept. 2017). “Challenges in Data-to-Document Generation”. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Ed. by Martha Palmer, Rebecca Hwa, and Sebastian Riedel. Copenhagen, Denmark: Association for Computational Linguistics, pp. 2253–2263. DOI: 10.18653/v1/D17-1239. URL: <https://aclanthology.org/D17-1239/>.