UNTERSUCHUNG DER EFFIZIENZ EINES ASSISTENZSYSTEMS FÜR TEXTANNOTATIONEN

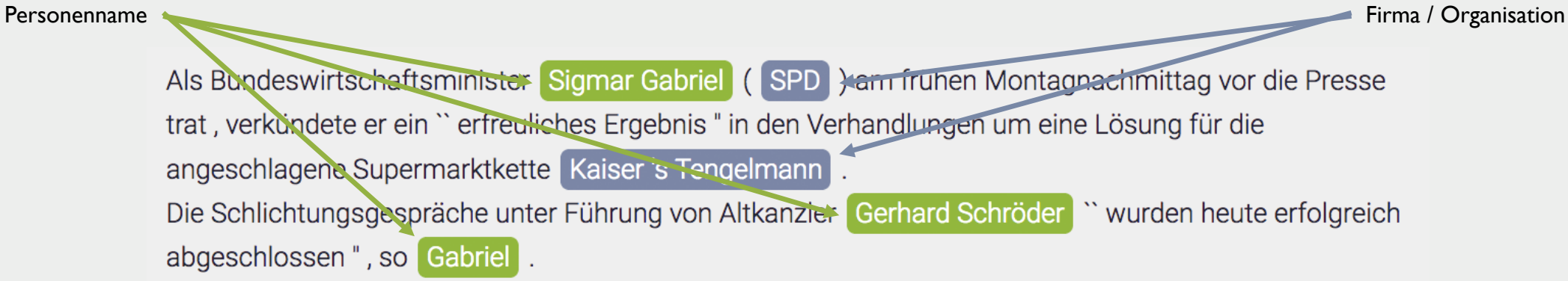
Bachelorarbeit von Robert Greinacher

MOTIVATION

Effizienz eines Assistenzsystems für Textannotationen

TEXTANNOTATIONEN

Einzelne oder mehrere Worte in Text zu markieren (um Eigenschaften zu kennzeichnen).



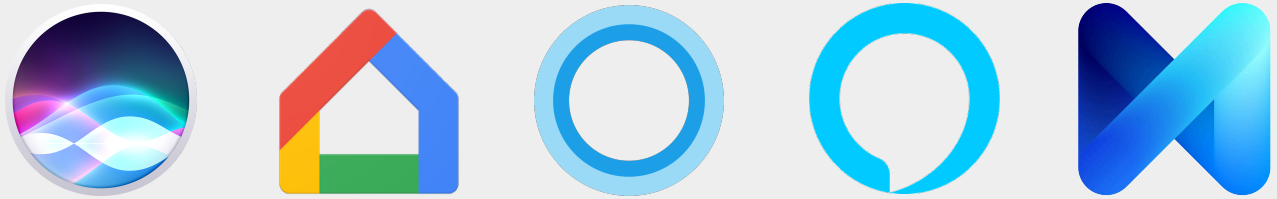
- Annotierte Texte als Trainingsdaten für *Machine Learning*

TEXTANNOTATIONEN FÜR MACHINE LEARNING

Zusammenhänge aus Text erfassen

- Text für Maschinen nur Zeichenfolgen
- ML kann diese Zeichenfolgen analysieren
- ➔ Textverständnis simulieren

- Beispiel:



TEXTANNOTATIONEN FÜR MACHINE LEARNING

Zusammenhänge aus Text erfassen

- Text für Maschinen nur Zeichenfolgen
- ML kann diese Zeichenfolgen analysieren
- ➔ Textverständnis simulieren

“Schreib Vera eine SMS wie das Wetter in Berlin gerade ist.“

- Eigennamen erkennen:

  Um Wen oder Was geht es?

TEXTANNOTATIONEN FÜR MACHINE LEARNING

Zusammenhänge aus Text erfassen

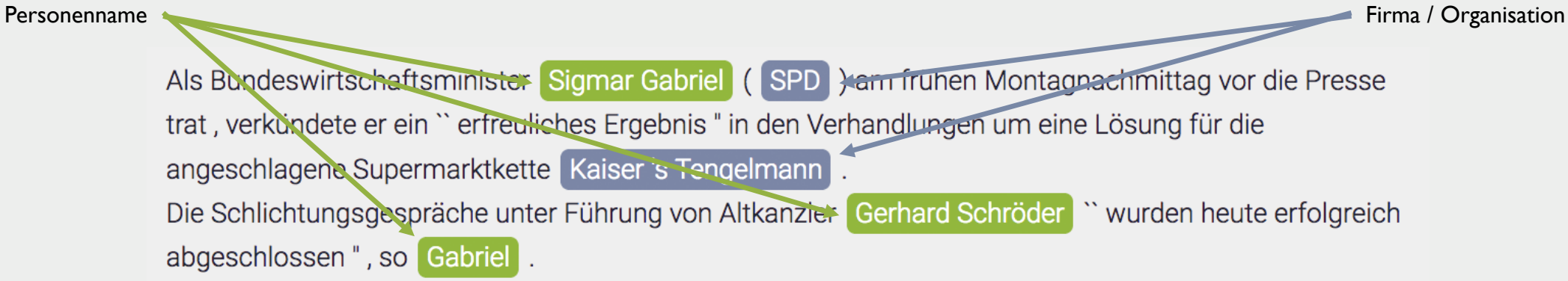
- Text für Maschinen nur Zeichenfolgen
- ML kann diese Zeichenfolgen analysieren
- ➔ Textverständnis simulieren

- Eigennamen erkennen:

  Um Wen oder Was geht es?
- Satzstruktur erkennen
- Schlüsselworte erkennen

TEXTANNOTATIONEN

Einzelne oder mehrere Worte in Text zu markieren (um Eigenschaften zu kennzeichnen).



- Annotierte Texte als Trainingsdaten für *Machine Learning*
- Ziel hier: Modell zur Erkennung von Personennamen und Firmen- bzw. Organisationsnamen

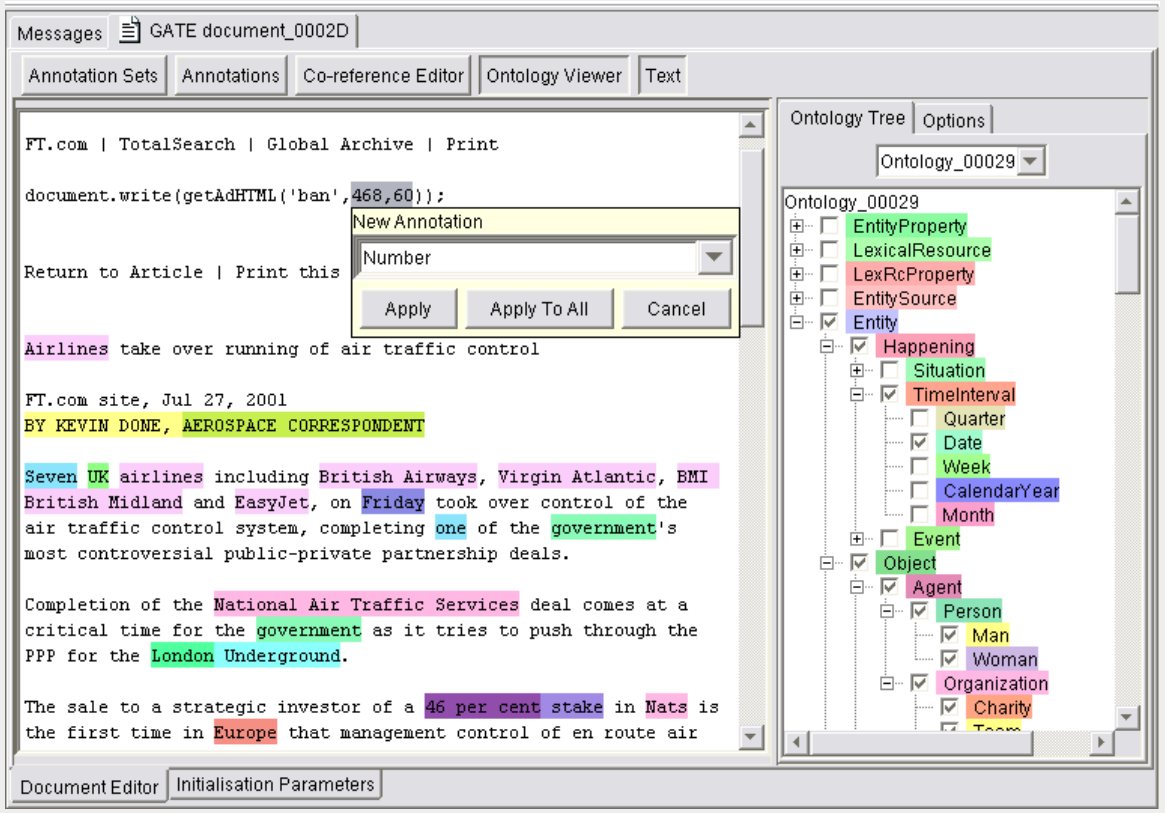
TEXTANNOTATIONEN

Bisherige Systeme für Textannotationen sind ungenügend

- Aufgabe generell sehr monoton & beanspruchend
- Sehr langwierig, dadurch teuer
- Interfaces meist unintuitiv, ggf. viel Vorkenntnisse notwendig

Wie machen wir Textannotationen einfacher?

- Neues Interface (Schlicht / stark Use Case orientiert)
- Sich iterativ verbessernde Assistenz um Belastung zu minimieren

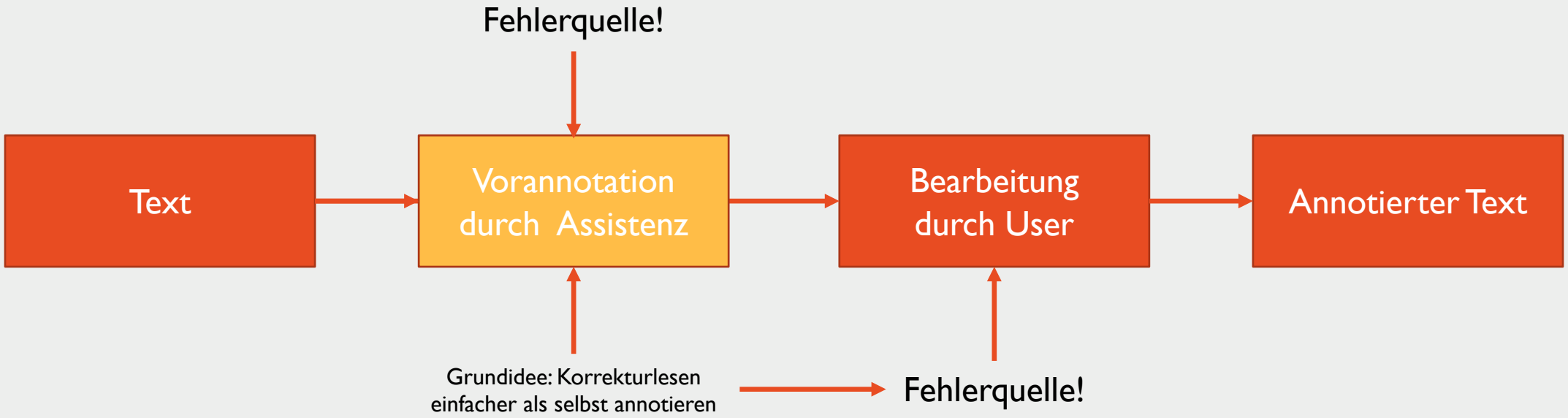


- GATE, gate.ac.uk

MOTIVATION

Effizienz eines Assistenzsystems für Textannotationen ✔

ASSISTENZSYSTEM



ASSISTENZSYSTEM

- Generiert Vorannotationen
  - Grundidee: Korrekturlesen ist einfacher als selbst annotieren
- Lernt iterativ
  - Lernt von bereits gemachten Annotationen
  - Bessere Vorschläge über die Zeit

Beispiele:

ohne Assistenz:

Als Bundeswirtschaftsminister Sigmar Gabriel (SPD) am frühen Montagnachmittag vor die Presse trat , verkündete er ein

korrekte Vorannotation der Assistenz:

Als Bundeswirtschaftsminister Sigmar Gabriel (SPD) am frühen Montagnachmittag vor die Presse trat , verkündete er ein

ASSISTENZSYSTEM

- Kann Fehler machen
  - Die von den Usern korrigiert werden müssen
- In dieser Untersuchung
  - Drei unterschiedliche Leistungsstufen der Assistenz:
    - 10% richtige Vorschläge
    - 50% richtige Vorschläge
    - 90% richtige Vorschläge
  - Konstante Leistung pro VP (Simulation)

Beispiele:

ohne Assistenz:

Als Bundeswirtschaftsminister Sigmar Gabriel (SPD) am frühen Montagnachmittag vor die Presse trat , verkündete er ein

korrekte Vorannotation der Assistenz:

Als Bundeswirtschaftsminister Sigmar Gabriel (SPD) am frühen Montagnachmittag vor die Presse trat , verkündete er ein

fehlerhafte Vorannotation der Assistenz:

Als Bundeswirtschaftsminister Sigmar Gabriel (SPD) am frühen Montagnachmittag vor die Presse trat , verkündete er ein

MOTIVATION

Effizienz eines Assistenzsystems für Textannotationen

✔ ✔

EFFIZIENZ

"Effizient arbeiten bedeutet, so zu arbeiten, dass erzieltes Ergebnis und eingesetzte Mittel in einem möglichst günstigen Kosten-Nutzen-Verhältnis stehen und der Nutzen dabei größer ist als die Kosten."

- Wikipedia

- Richtigkeit
  - Werden mit Assistenz mehr Annotationen richtig gemacht als ohne Assistenz?
- Tempo
  - Werden die Annotationen mit Assistenz schneller gemacht als ohne Assistenz?
- Übersehene Annotationsstellen
  - Werden mit Assistenz weniger Annotationsstellen übersehen als ohne?
- Zugabe: persönliche Empfindungen
  - Verändert sich die empfundene Beanspruchung und Monotonie mit einer Assistenz?

HYPOTHESEN

HYPOTHESEN

	Richtigkeit	Tempo	Übersehene Annotationsstellen
10% richtige Assistenz			
50% richtige Assistenz			
90% richtige Assistenz			

- "Grundidee: Korrekturlesen (und korrigieren) ist einfacher als selbst annotieren."

HYPOTHESEN

	Richtigkeit	Tempo	Übersehene Annotationsstellen
10% richtige Assistenz	mehr richtig als ohne		
50% richtige Assistenz	mehr richtig als ohne		
90% richtige Assistenz	mehr richtig als ohne		

- "Grundidee: Korrekturlesen (und korrigieren) ist einfacher als selbst annotieren."

HYPOTHESEN

	Richtigkeit	Tempo	Übersehene Annotationsstellen
10% richtige Assistenz	mehr richtig als ohne	schneller als ohne	
50% richtige Assistenz	mehr richtig als ohne	schneller als ohne	
90% richtige Assistenz	mehr richtig als ohne	schneller als ohne	

- "Grundidee: Korrekturlesen (und korrigieren) ist einfacher als selbst annotieren."

HYPOTHESEN

	Richtigkeit	Tempo	Übersehene Annotationsstellen
10% richtige Assistenz	mehr richtig als ohne	schneller als ohne	weniger übersehen als ohne
50% richtige Assistenz	mehr richtig als ohne	schneller als ohne	weniger übersehen als ohne
90% richtige Assistenz	mehr richtig als ohne	schneller als ohne	weniger übersehen als ohne

- "Grundidee: Korrekturlesen (und korrigieren) ist einfacher als selbst annotieren."
- ➔ Jede Assistenz macht die Annotationsaufgabe besser als keine Assistenz

HYPOTHESEN

	Richtigkeit	Tempo	Übersehene Annotationsstellen
10% richtige Assistenz	mehr richtig als ohne	schneller als ohne	weniger übersehen als ohne
50% richtige Assistenz	mehr richtig als ohne	schneller als ohne	weniger übersehen als ohne
90% richtige Assistenz	mehr richtig als ohne	schneller als ohne	weniger übersehen als ohne
10% < 50%	50% richtige Assistenz macht noch mehr richtig	50% richtige Assistenz noch schneller	50% richtige Assistenz noch weniger übersehen
50% < 90%	90% richtige Assistenz macht noch mehr richtig	90% richtige Assistenz noch schneller	90% richtige Assistenz noch weniger übersehen

- ➔ Der Einfluss des Assistenzsystems nimmt proportional zur Richtigkeit des Assistenzsystems zu

VERSUCHSDESIGN

Aufgabe und Versuchsaufbau

BEARBEITUNGSGEGENSTAND

Als Bundeswirtschaftsminister Sigmar Gabriel (SPD) am frühen Montagnachmittag vor die Presse trat , verkündete er ein `` erfreuliches Ergebnis '' in den Verhandlungen um eine Lösung für die angeschlagene Supermarktkette Kaiser 's Tengelmann .
Die Schlichtungsgespräche unter Führung von Altkanzler Gerhard Schröder `` wurden heute erfolgreich abgeschlossen '' , so Gabriel .

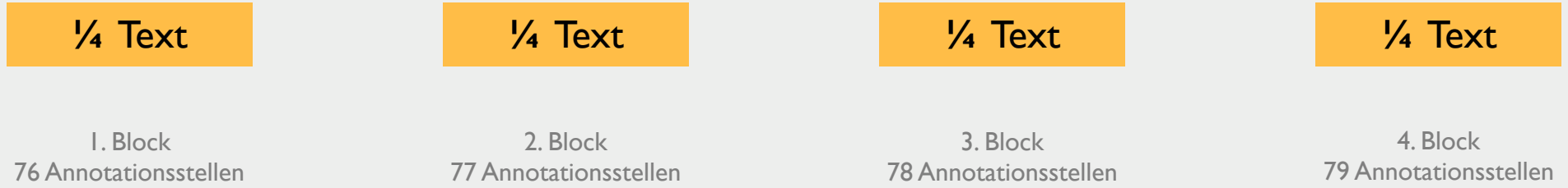
- 14 Nachrichtentexte verschiedener Themen
- 5989 Worte, etwa 25 min Lesezeit
- 73 Absätze, 305 Sätze
- Ausgewählt nach Länge und Anzahl der Annotationsstellen

- Annotation von Personen- und Organisationsnamen
- Insgesamt 310 Annotationsstellen

VERSUCHSAUFBAU

Text

VERSUCHSAUFBAU



- Text absatzweise auf vier Blöcke verteilt
  - möglichst gleich viele Annotationsstellen pro Block

VERSUCHSAUFBAU

Assistenz	¬ Assistenz	Assistenz	¬ Assistenz

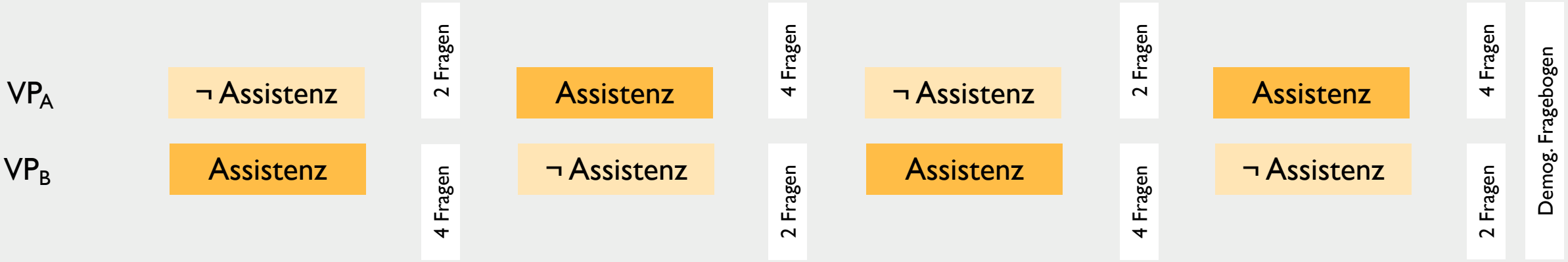
- Text absatzweise auf vier Blöcke verteilt
- Zwei Blöcke mit Assistenz, zwei ohne (Messzeitpunkt, within Faktor)

VERSUCHSAUFBAU

VP_A	¬ Assistenz	Assistenz	¬ Assistenz	Assistenz
VP_B	Assistenz	¬ Assistenz	Assistenz	¬ Assistenz

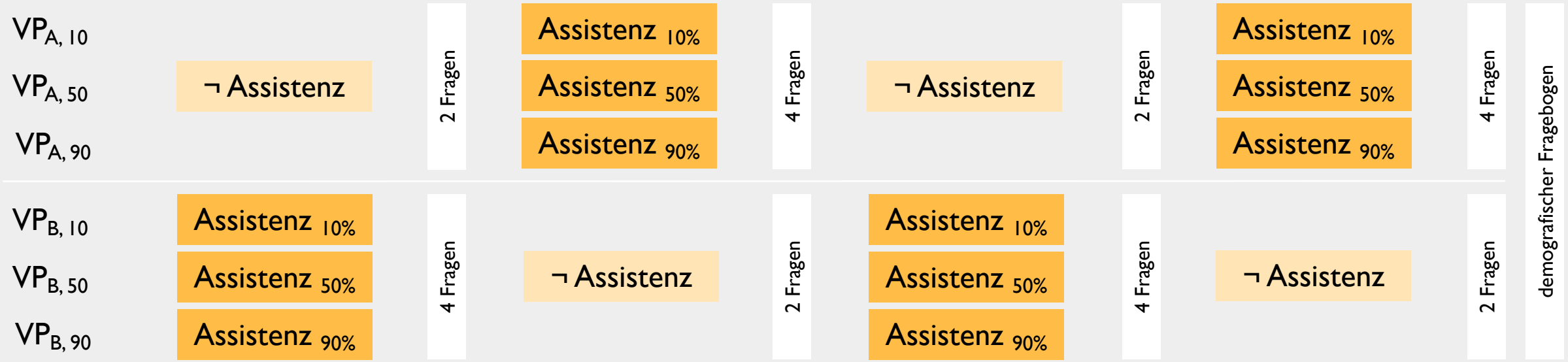
- Text absatzweise auf vier Blöcke verteilt
- Zwei Blöcke mit Assistenz, zwei ohne (Messzeitpunkt, within Faktor)
- Reihenfolge alterniert zwischen VPs (between Faktor, ausbalanciert)

VERSUCHSAUFBAU



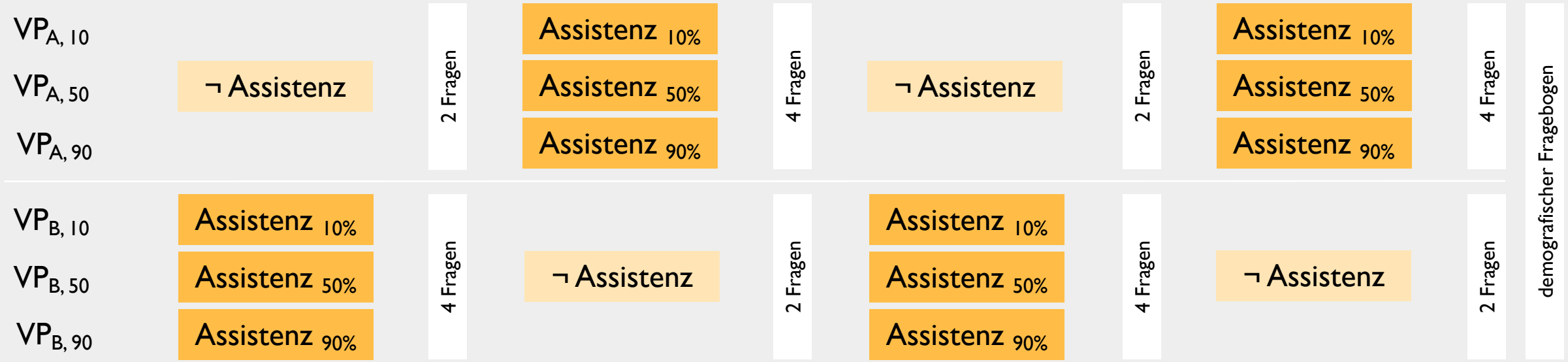
- Text absatzweise auf vier Blöcke verteilt
- Zwei Blöcke mit Assistenz, zwei ohne (Messzeitpunkt, within Faktor)
- Reihenfolge alterniert zwischen VPs (between Faktor, ausbalanciert)
- Zwei bzw. vier Fragen nach jedem Block, demografischer Fragebogen zum Ende

VERSUCHSAUFBAU



- Text absatzweise auf vier Blöcke verteilt
- Zwei Blöcke mit Assistenz, zwei ohne (Messzeitpunkt, within Faktor)
- Reihenfolge alterniert zwischen VPs (between Faktor, ausbalanciert)
- Zwei bzw. vier Fragen nach jedem Block, demografischer Fragebogen zum Ende
- 3 Stufen der Assistenz: 10% korrekt / 50% korrekt / 90% korrekt (between Faktor)

VERSUCHSAUFBAU

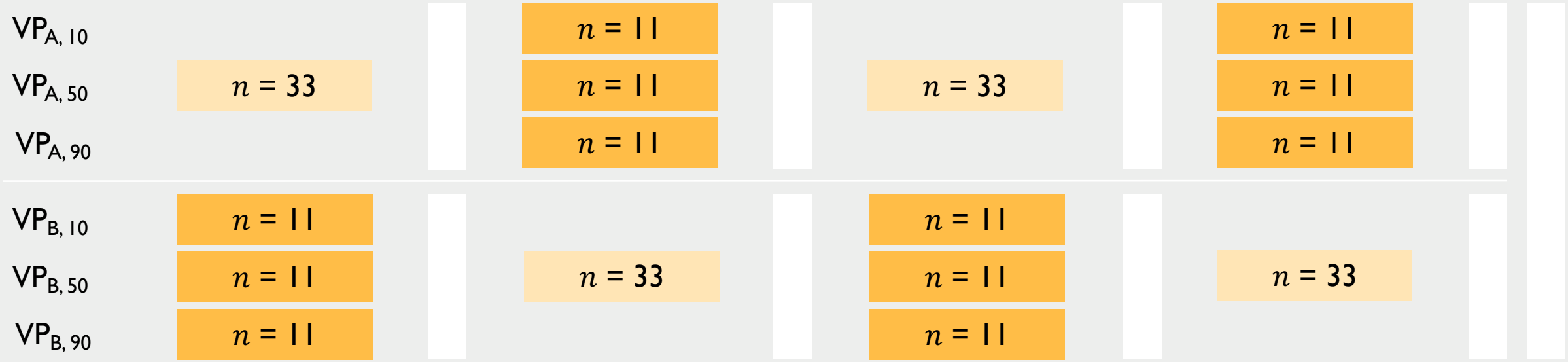


A priori Power Analyse (One Way ANOVA):

- Drei Gruppen
- Angenommene Effektgröße f = 0,4 / α = 0,05 / Power = 0,8

➔ N = 66

VERSUCHSAUFBAU



UV:

- Stufe des Assistenzsystems (between)
  - 3 Stufen: 10% korrekt / 50% korrekt / 90% korrekt
- Messzeitpunkt (within)
  - 2 Stufen: erste Hälfte / zweite Hälfte

➔ 2 x 3 Design

AV:

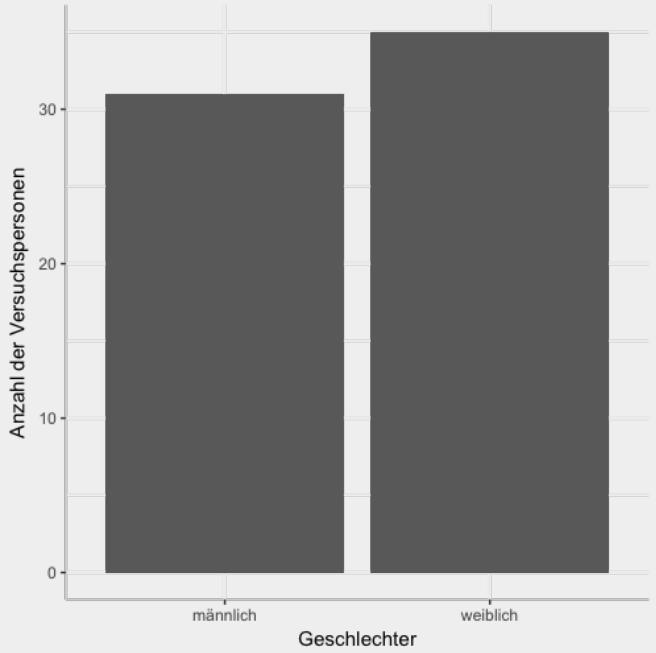
- Bearbeitungszeit pro Absatz
  - Durchschnittliche Zeit pro Annotation pro Block
- Annotation pro Annotationsstelle
  - Anzahl richtiger Annotationen pro Block
  - Anzahl übersehene Annotationen pro Block

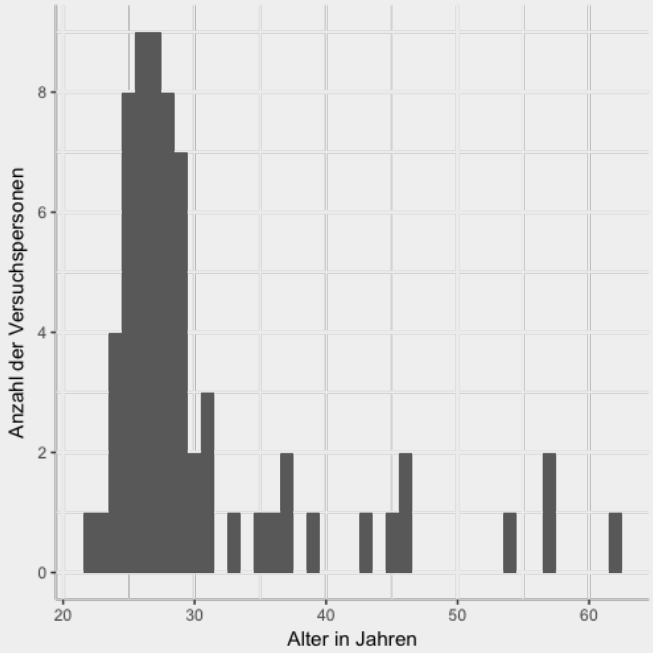
VERSUCHSDURCHFÜHRUNG

- Laborbedingungen
- Durchführung zwischen 20. Februar und 17. März

- 35 weibliche, 31 männliche VP
- Schnitt: 30,68 Jahre (SD: 8,68 Jahre)

- Incentivierung
  - Verlosungsteilnahme zweier Gutscheine
  - VP Stunde
  - 10€ Bargeld





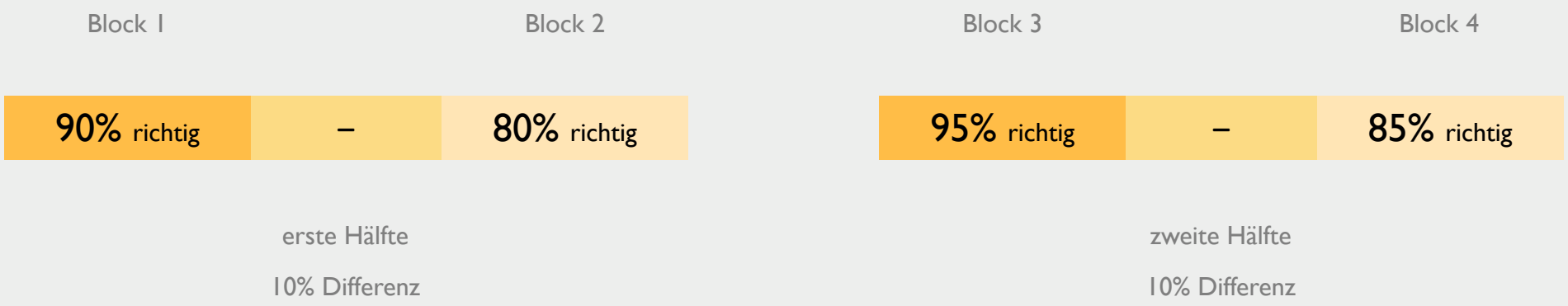
DATEN

Auswertung und Ergebnisse

AUSWERTUNG

- Zwei Blöcke mit Assistenz, zwei ohne (Messzeitpunkt, within Faktor)
  - Differenz zwischen Baseline und Manipulation
  - Block mit Assistenz Minus Block ohne Assistenz

AUSWERTUNG



Beispiel für eine VP

- Erste Hälfte: 10% Differenz
- Zweite Hälfte: 10% Differenz
- ➔ Die VP hat mit Assistenz **10% mehr Annotationen richtig** gemacht als ohne

AUSWERTUNG: HYPOTHESEN 1-3 (RICHTIGKEIT)

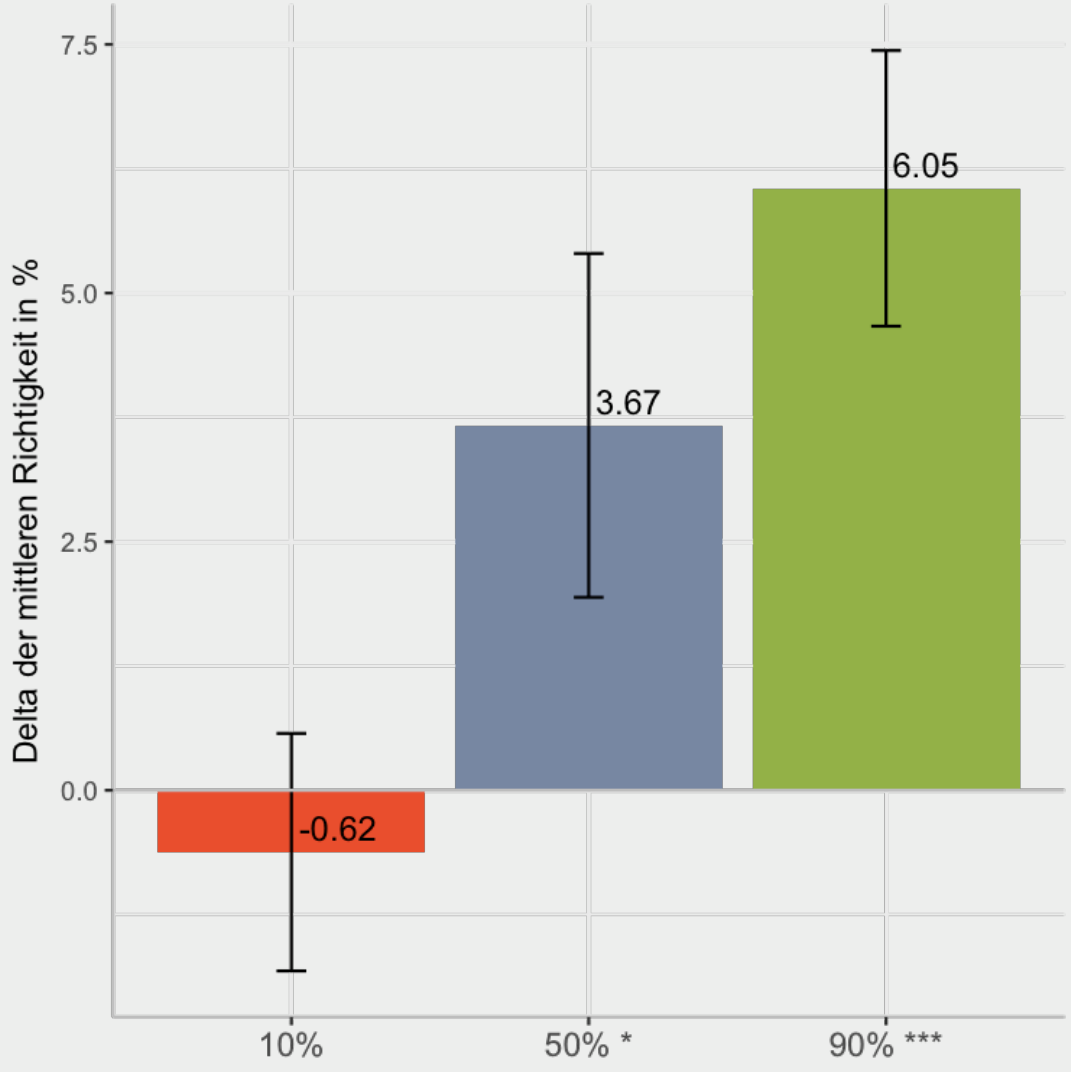
Stufe der Assistenz	Hypothese	Test	n	Mean	t	p	Signifikant?
10% richtige Assistenz	mehr richtig als ohne	One Sample T-Test	Je 22	-0,6226	-0.5206	0,6081	✗
50% richtige Assistenz	mehr richtig als ohne			3,6695	2.1231	0,0458	✓
90% richtige Assistenz	mehr richtig als ohne			6,0529	4.3667	0,0003	✓

- Signifikanzniveau: 0,05
- Anmerkungen zum Durchschnitt
  - Negative Differenz: Das Assistenzsystem wirkt **senkend** auf die Richtigkeit (vgl. 10%)
  - Positive Differenz: Das Assistenzsystem wirkt **steigernd** auf die Richtigkeit (vgl. 50% / 90%)
- ➔ Die Assistenz unterstützt in den Stufen 50% und 90%

AUSWERTUNG: HYPOTHESEN 1-3 (RICHTIGKEIT)

Stufe der Assistenz	Hypothese
10% richtige Assistenz	mehr richtig als ohne
50% richtige Assistenz	mehr richtig als ohne
90% richtige Assistenz	mehr richtig als ohne

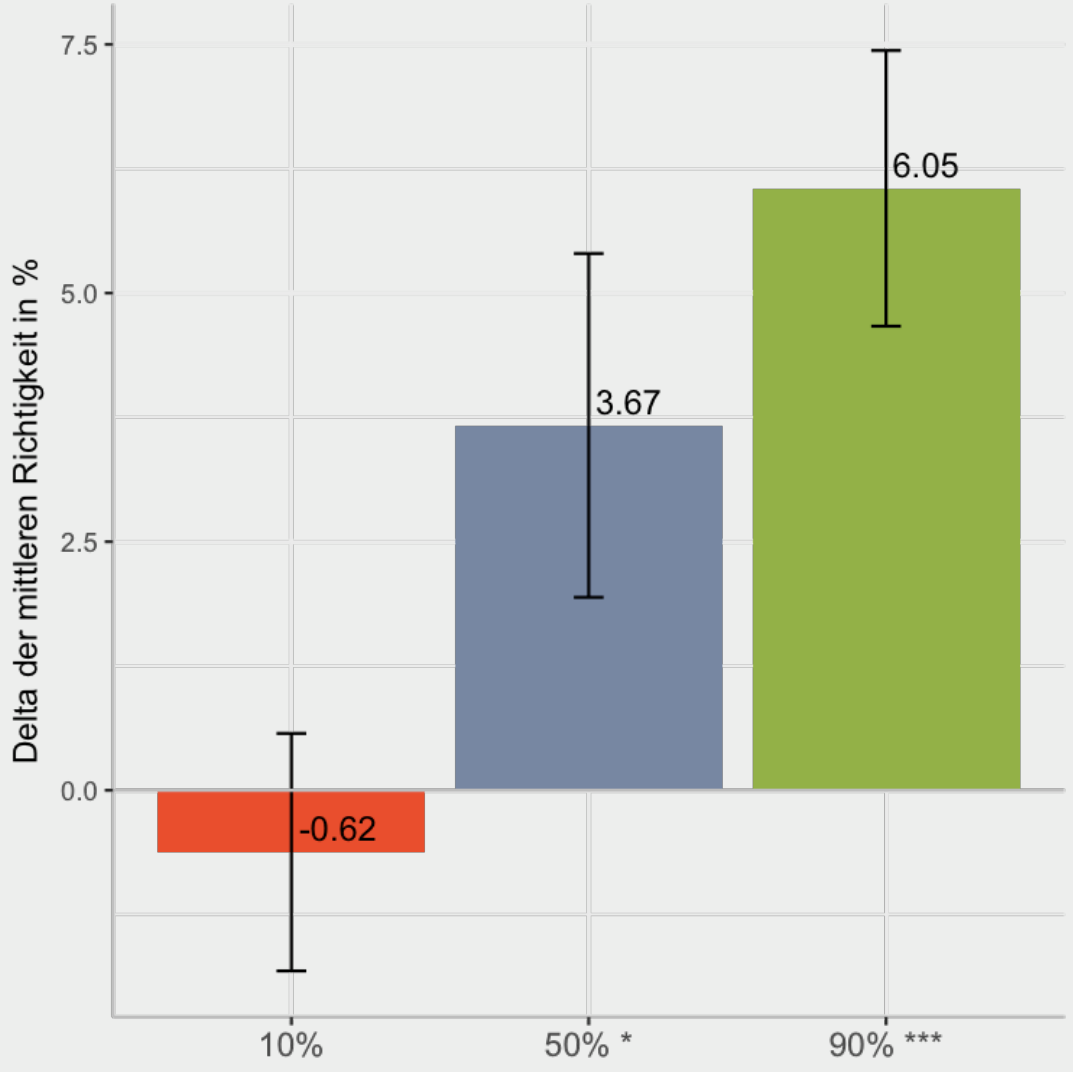
- Baseline: 83,93% richtige Annotationen
- 3x One Sample T-Test



AUSWERTUNG: HYPOTHESEN 1-3 (RICHTIGKEIT)

Stufe der Assistenz	Hypothese	Signifikant? α = 0,05
10% richtige Assistenz	mehr richtig als ohne	✗
50% richtige Assistenz	mehr richtig als ohne	✓
90% richtige Assistenz	mehr richtig als ohne	✓

- Baseline: 83,93% richtige Annotationen
- 3x One Sample T-Test
- ➔ Die Assistenz unterstützt in den Stufen 50% und 90%
- r = 0.38

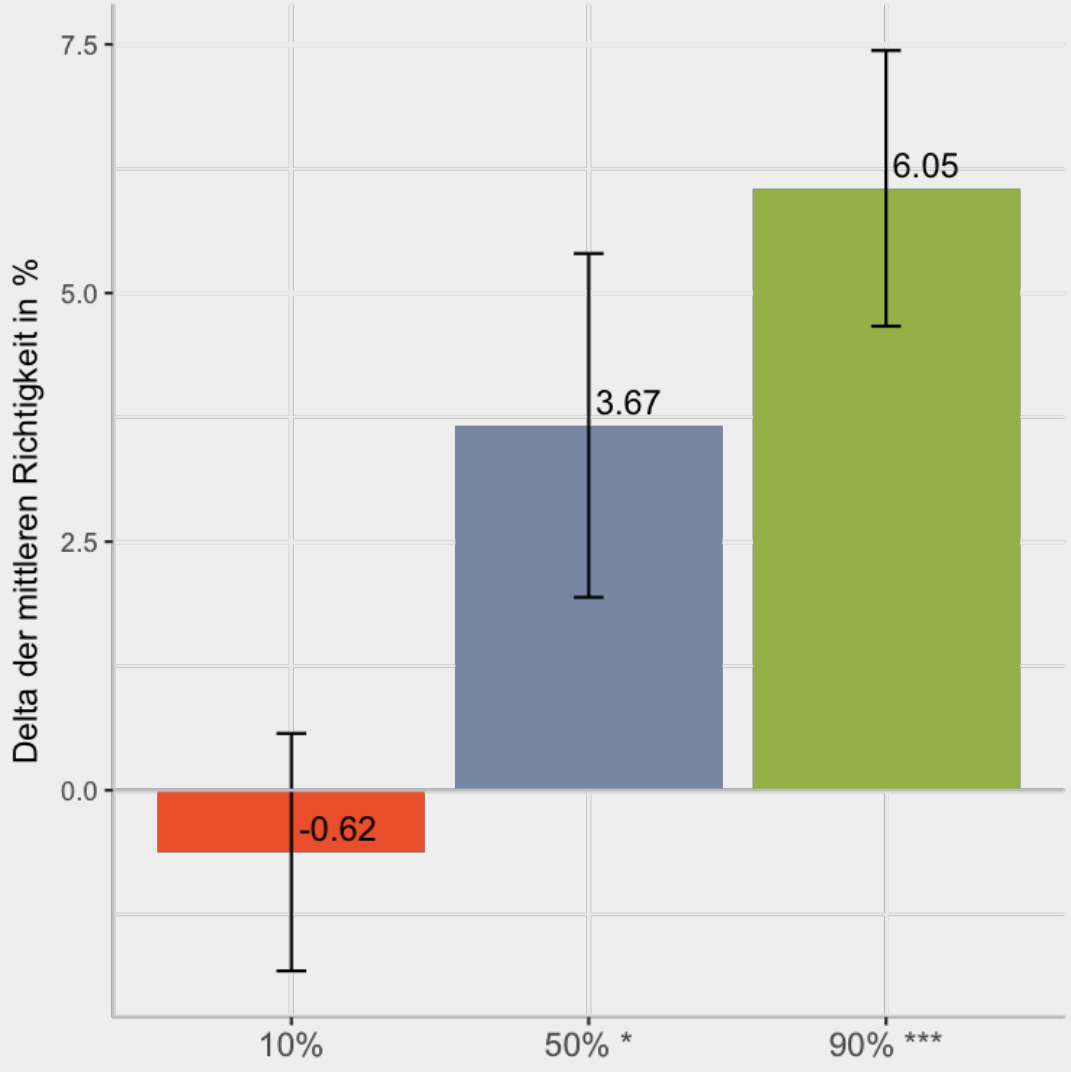


AUSWERTUNG: HYPOTHESEN 4 & 5 (RICHTIGKEIT)

Hypothese * bzgl. der Richtigkeit	Signifikant?
10% < 50%*	
50% < 90%*	

- Typ-3 ANOVA (Stufe des Assistenzsystems × Block)
- Haupteffekt der Stufe des Assistenzsystems
- kein signifikanter Haupteffekt des Blocks
- kein signifikanter Interaktionseffekt

- T-Test: 10% vs. 50% und 50% vs. 90%

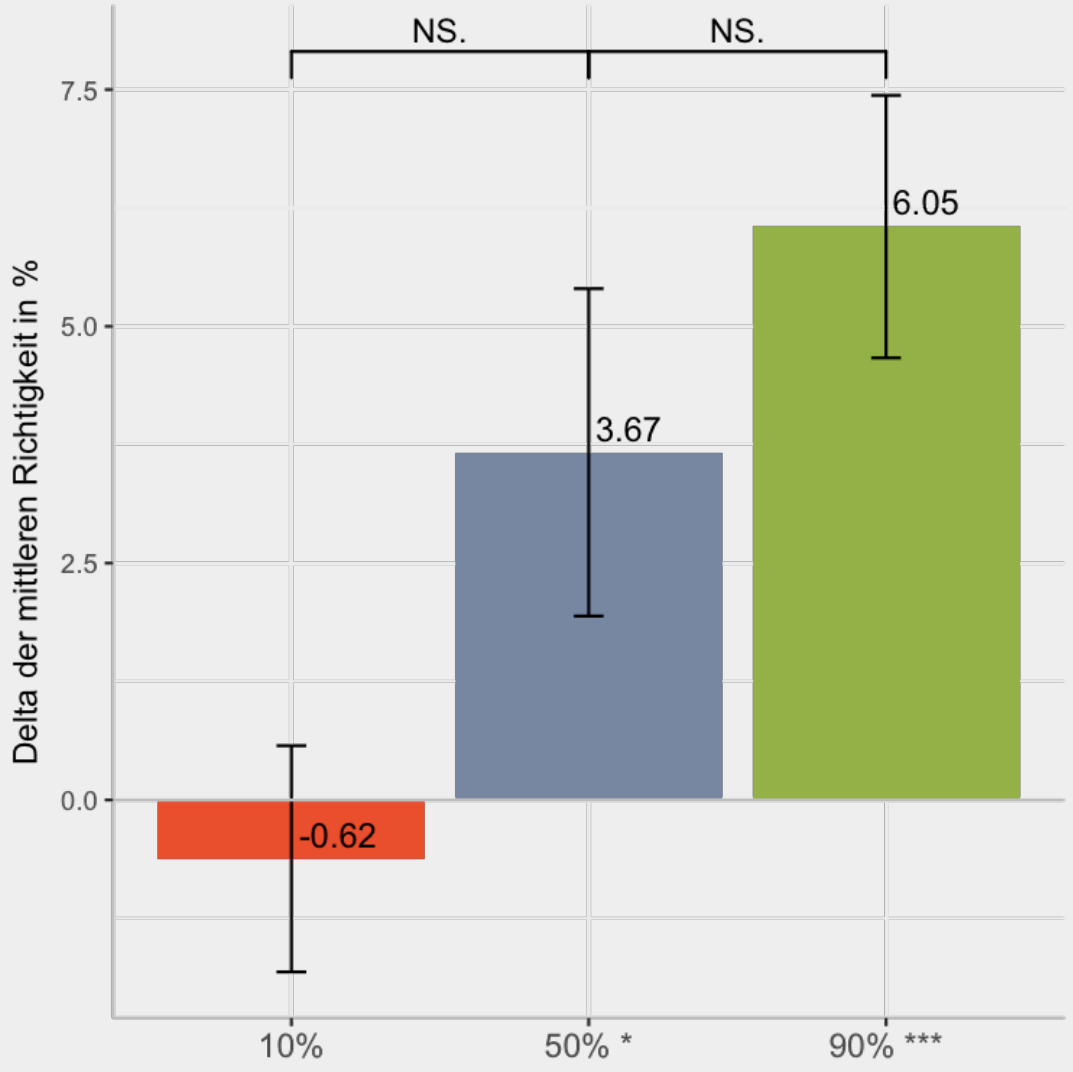


AUSWERTUNG: HYPOTHESEN 4 & 5 (RICHTIGKEIT)

Hypothese * bzgl. der Richtigkeit	Signifikant? α = 0,025
10% < 50%*	✗
50% < 90%*	✗

- Typ-3 ANOVA (Stufe des Assistenzsystems × Block)
- Haupteffekt der Stufe des Assistenzsystems
- kein signifikanter Haupteffekt des Blocks
- kein signifikanter Interaktionseffekt

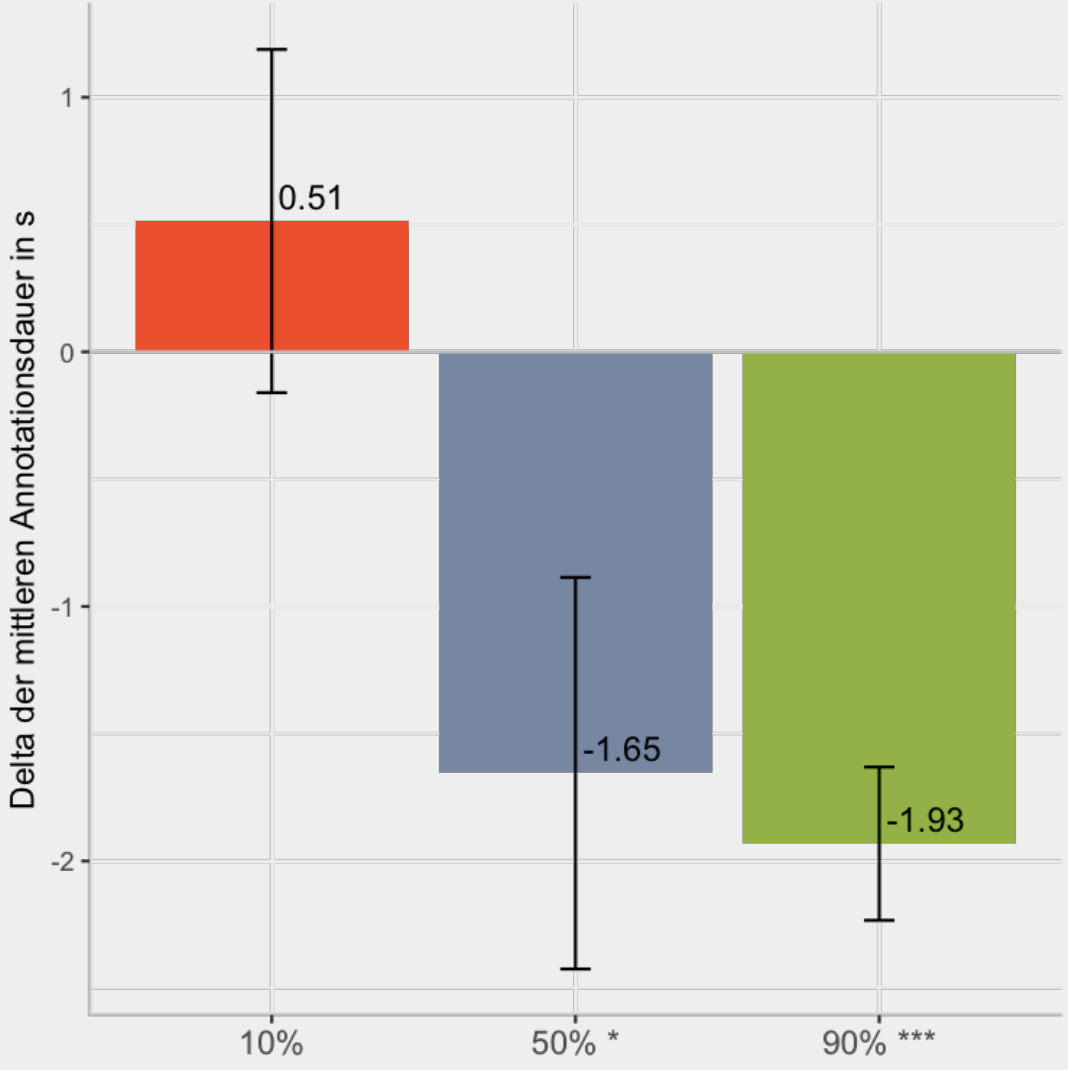
- T-Test: 10% vs. 50% und 50% vs. 90%
- ➔ Kein sign. Unterschied zur jeweils benachbarten Stufe



AUSWERTUNG: HYPOTHESEN 6-8 (TEMPO)

Stufe der Assistenz	Hypothese
10% richtige Assistenz	schneller als ohne
50% richtige Assistenz	schneller als ohne
90% richtige Assistenz	schneller als ohne

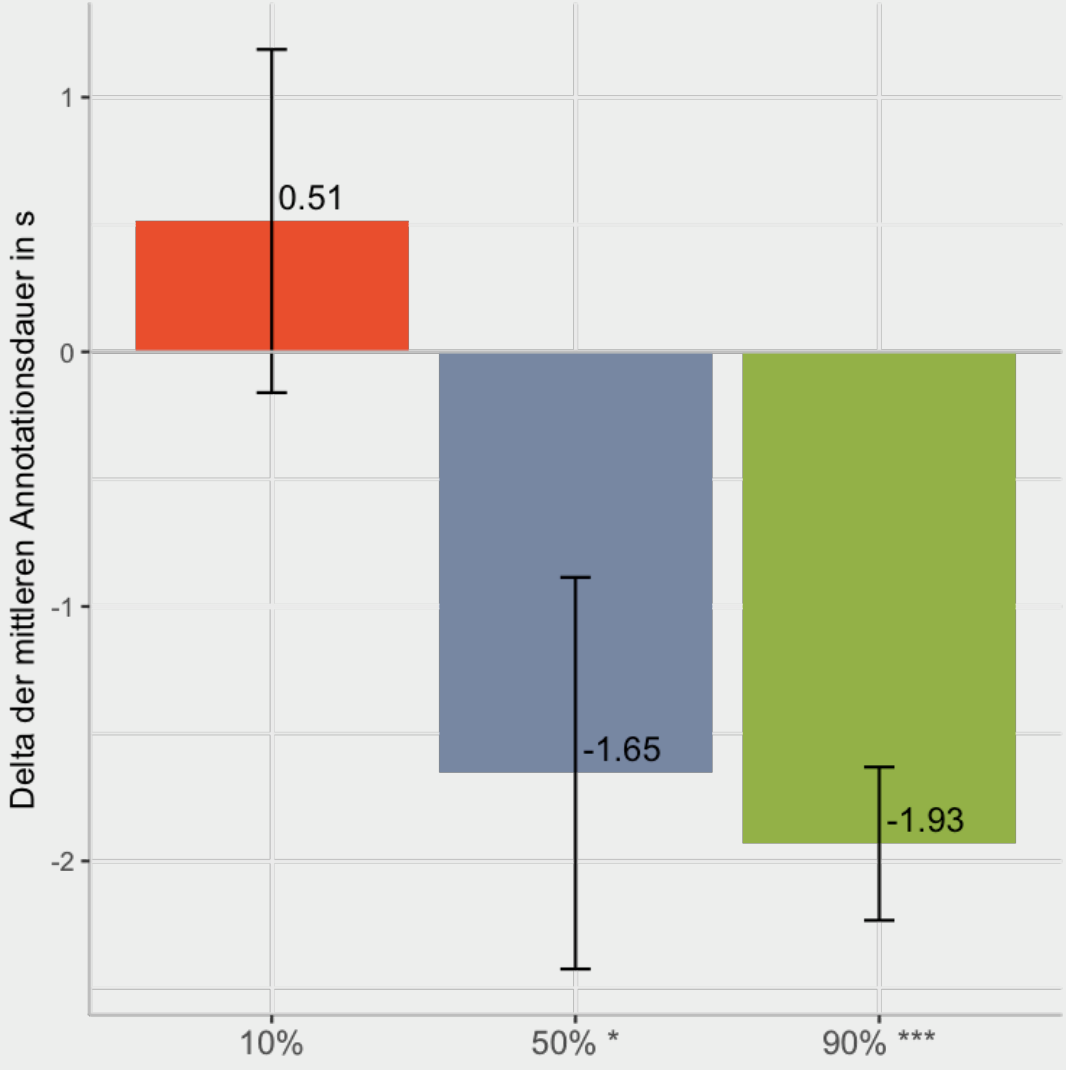
- Baseline: 8,19s pro Annotation
- 3x One Sample T-Test



AUSWERTUNG: HYPOTHESEN 6-8 (TEMPO)

Stufe der Assistenz	Hypothese	Signifikant? α = 0,05
10% richtige Assistenz	schneller als ohne	✗
50% richtige Assistenz	schneller als ohne	✓
90% richtige Assistenz	schneller als ohne	✓

- Baseline: 8,19s pro Annotation
- 3x One Sample T-Test
- ➔ Die Assistenz unterstützt in den Stufen 50% und 90%
- r = -0.33

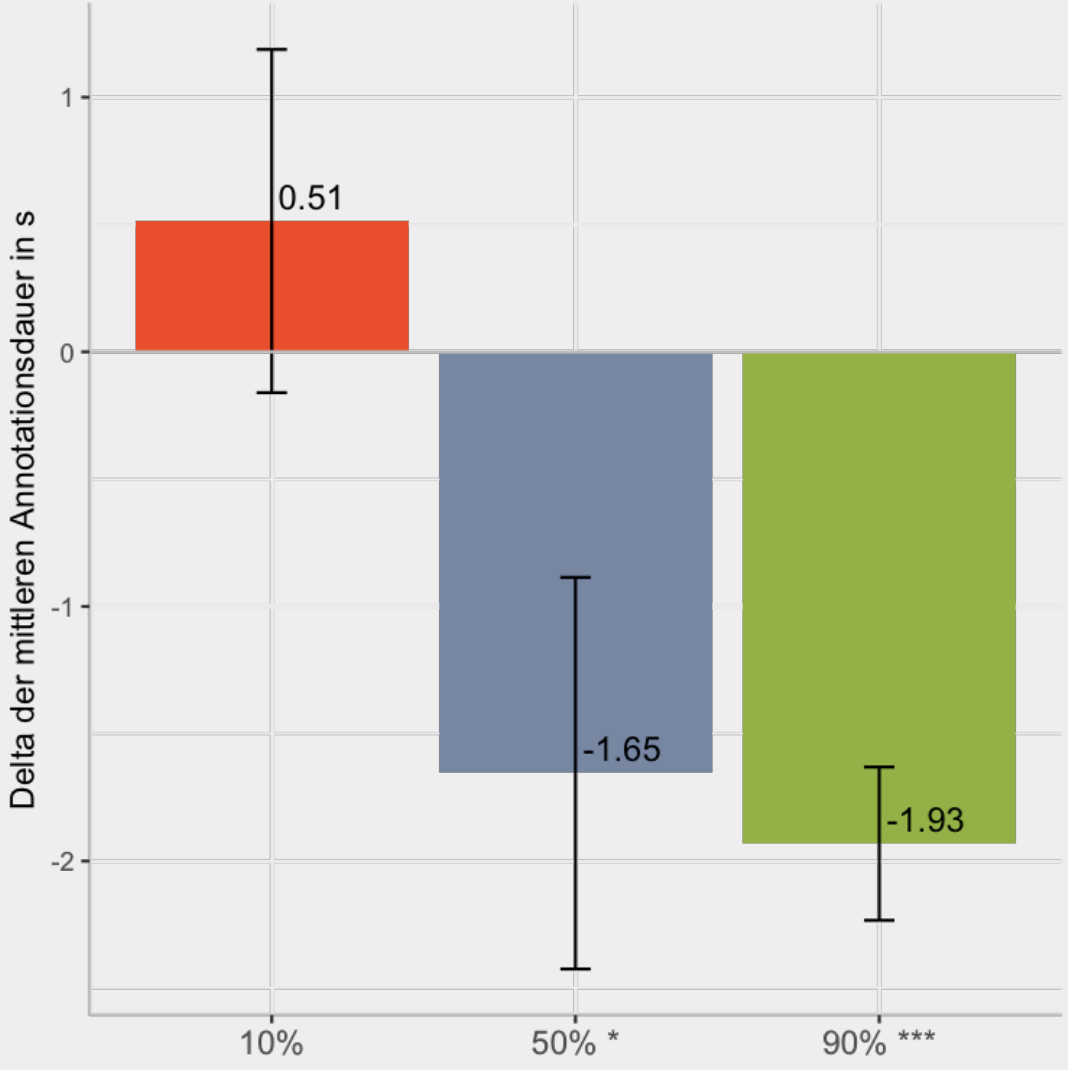


AUSWERTUNG: HYPOTHESEN 9 & 10 (TEMPO)

Hypothese * bzgl. des Tempos	Signifikant?
10% < 50%*	
50% < 90%*	

- Typ-3 ANOVA (Stufe des Assistenzsystems × Block)
- Haupteffekt der Stufe des Assistenzsystems
- kein signifikanter Haupteffekt des Blocks
- kein signifikanter Interaktionseffekt

- T-Test: 10% vs. 50% und 50% vs. 90%

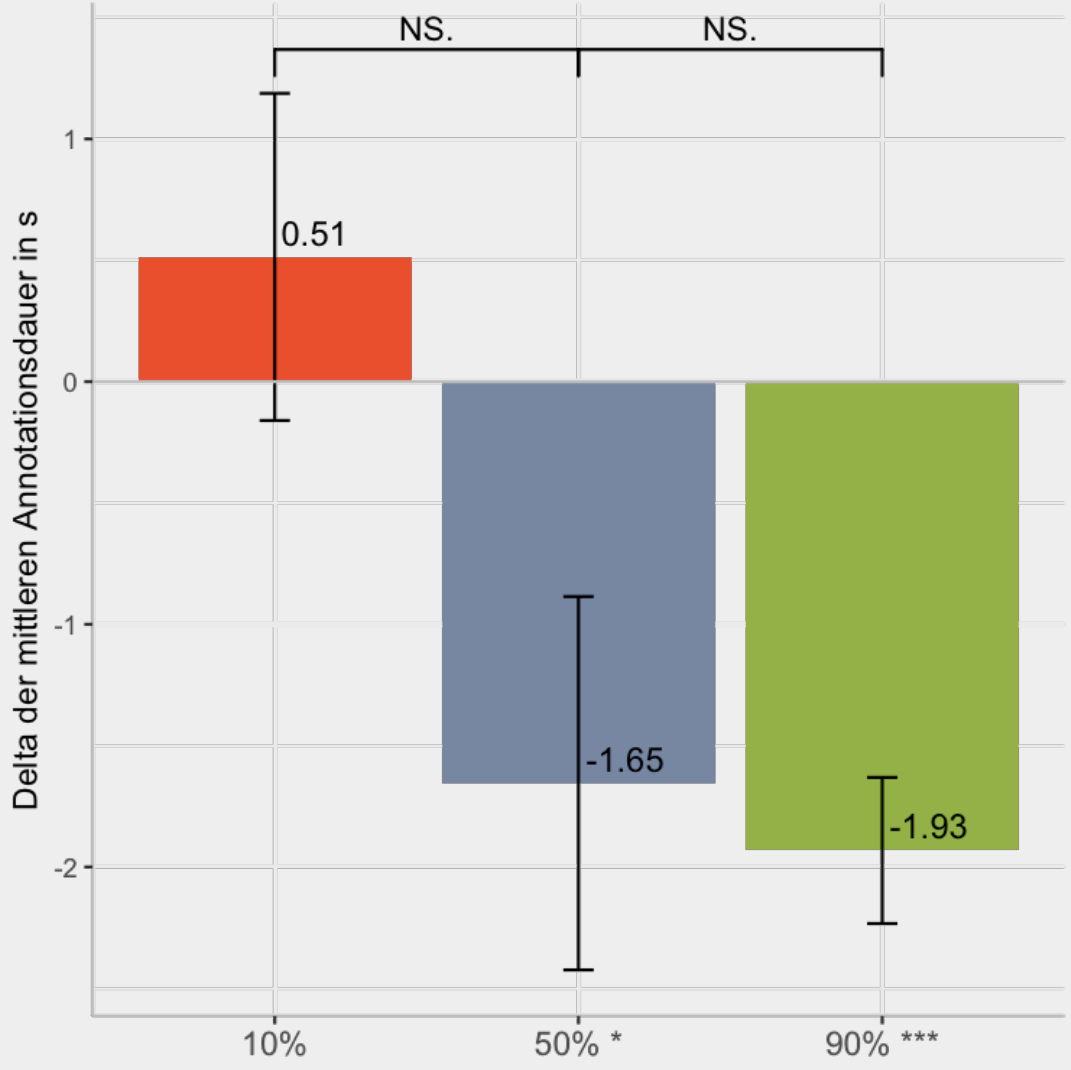


AUSWERTUNG: HYPOTHESEN 9 & 10 (TEMPO)

Hypothese * bzgl. des Tempos	Signifikant? α = 0,025
10% < 50%*	✗
50% < 90%*	✗

- Typ-3 ANOVA (Stufe des Assistenzsystems × Block)
- Haupteffekt der Stufe des Assistenzsystems
- kein signifikanter Haupteffekt des Blocks
- kein signifikanter Interaktionseffekt

- T-Test: 10% vs. 50% und 50% vs. 90%
- ➔ Kein sign. Unterschied zur jeweils benachbarten Stufe



AUSWERTUNG: HYPOTHESEN 7-9 (ÜBERSEHENE AS.)

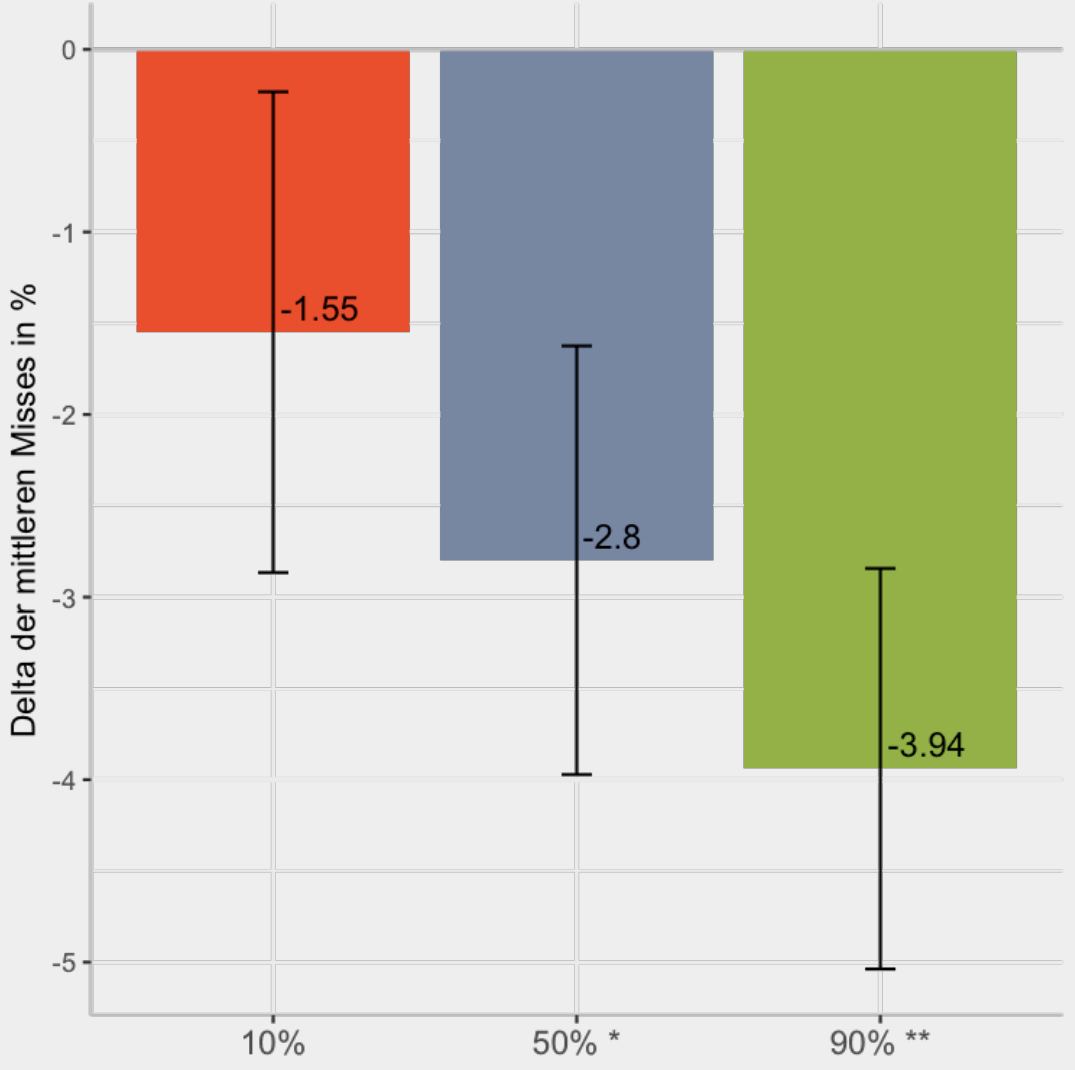
Stufe der Assistenz	Hypothese	Test	n	Mean	t	p	Signifikant?
10% richtige Assistenz	weniger übersehen als ohne	One Sample T-Test	Je 22	-0,0155	-1,1764	0,2526	✗
50% richtige Assistenz	weniger übersehen als ohne			-0,028	-2,3831	0,0267	✓
90% richtige Assistenz	weniger übersehen als ohne			-0,0394	-3.5905	0,0017	✓

- Signifikanzniveau: 0,05
- Anmerkungen zum Durchschnitt (Mean)
  - Negative Differenz: Das Assistenzsystem wirkt **senkend** auf die Zahl der übersehenen Annotationen
  - Positive Differenz: Das Assistenzsystem wirkt **steigernd** auf die Zahl der übersehenen Annotationen
- ➔ Die Assistenz unterstützt in den Stufen 50% und 90%

AUSWERTUNG: HYPOTHESEN 11-13 (ÜBERSEHENE AS.)

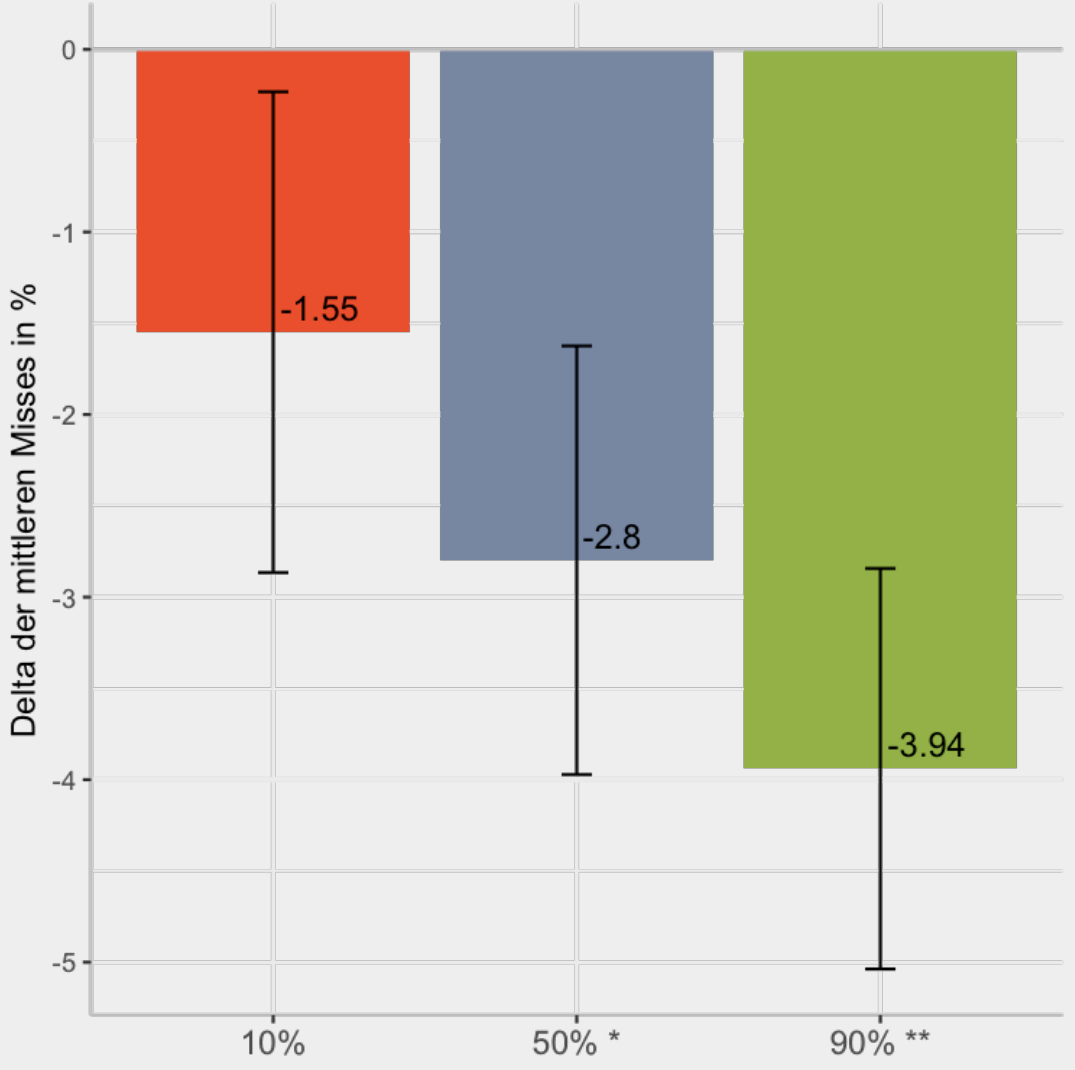
Stufe der Assistenz	Hypothese
10% richtige Assistenz	weniger übersehen als ohne
50% richtige Assistenz	weniger übersehen als ohne
90% richtige Assistenz	weniger übersehen als ohne

- Baseline: 7,69% übersehene Annotationen
- 3x One Sample T-Test



AUSWERTUNG: HYPOTHESEN 11-13 (ÜBERSEHENE AS.)

Stufe der Assistenz	Hypothese	Signifikant? α = 0,05
10% richtige Assistenz	weniger übersehen als ohne	✗
50% richtige Assistenz	weniger übersehen als ohne	✓
90% richtige Assistenz	weniger übersehen als ohne	✓



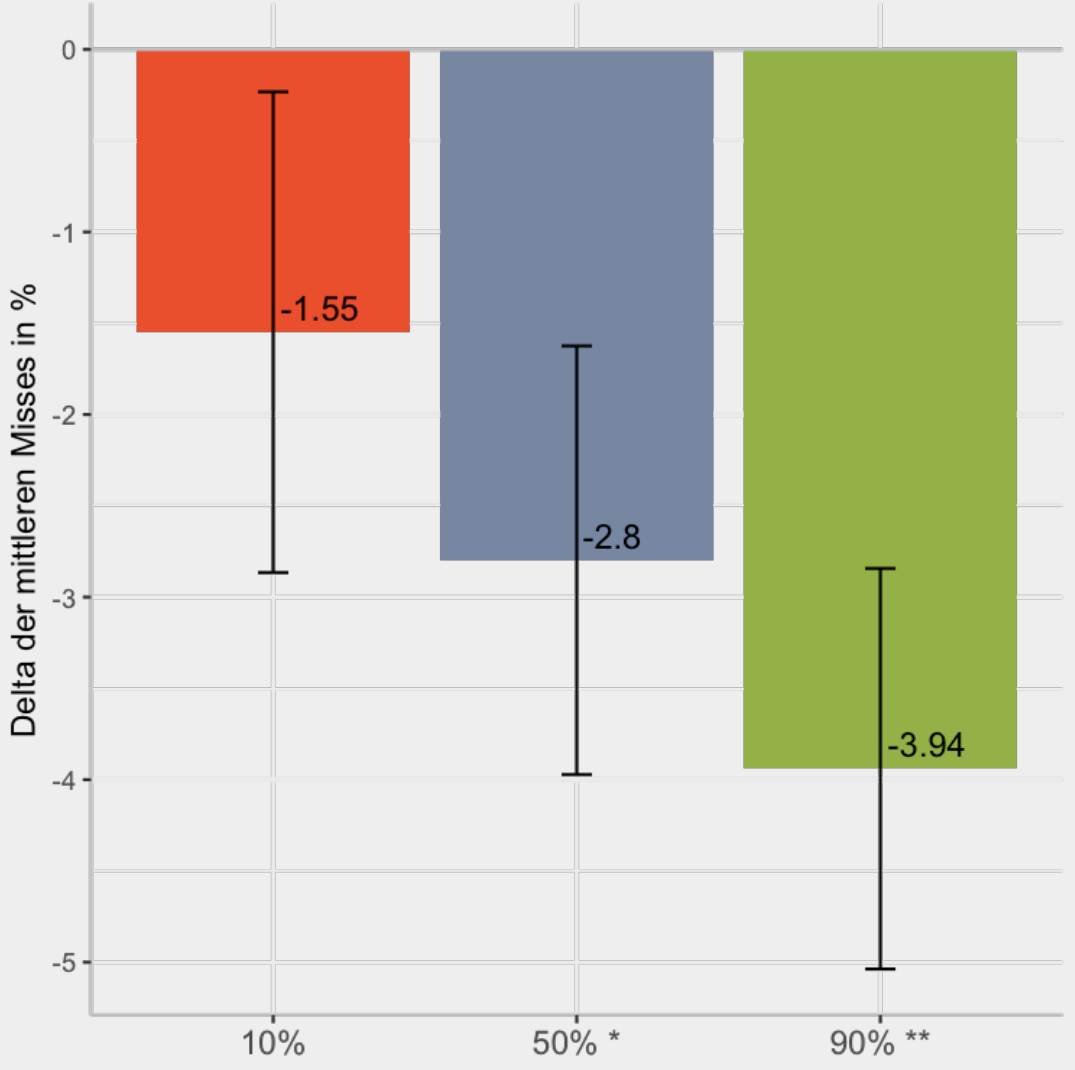
- Baseline: 7,69% übersehene Annotationen
- 3x One Sample T-Test
- ➔ Die Assistenz unterstützt in den Stufen 50% und 90%
- r = -0.17

AUSWERTUNG: HYPOTHESEN 14 & 15 (ÜBERSEHENE AS.)

Hypothese * bzgl. der übersehenen AS.	Signifikant?
10% < 50%*	
50% < 90%*	

- Typ-3 ANOVA (Stufe des Assistenzsystems × Block)
- kein signifikanter Haupteffekt der Stufe des Assistenzsystems
- Haupteffekt des Blocks
- kein signifikanter Interaktionseffekt

- T-Test: 10% vs. 50% und 50% vs. 90%

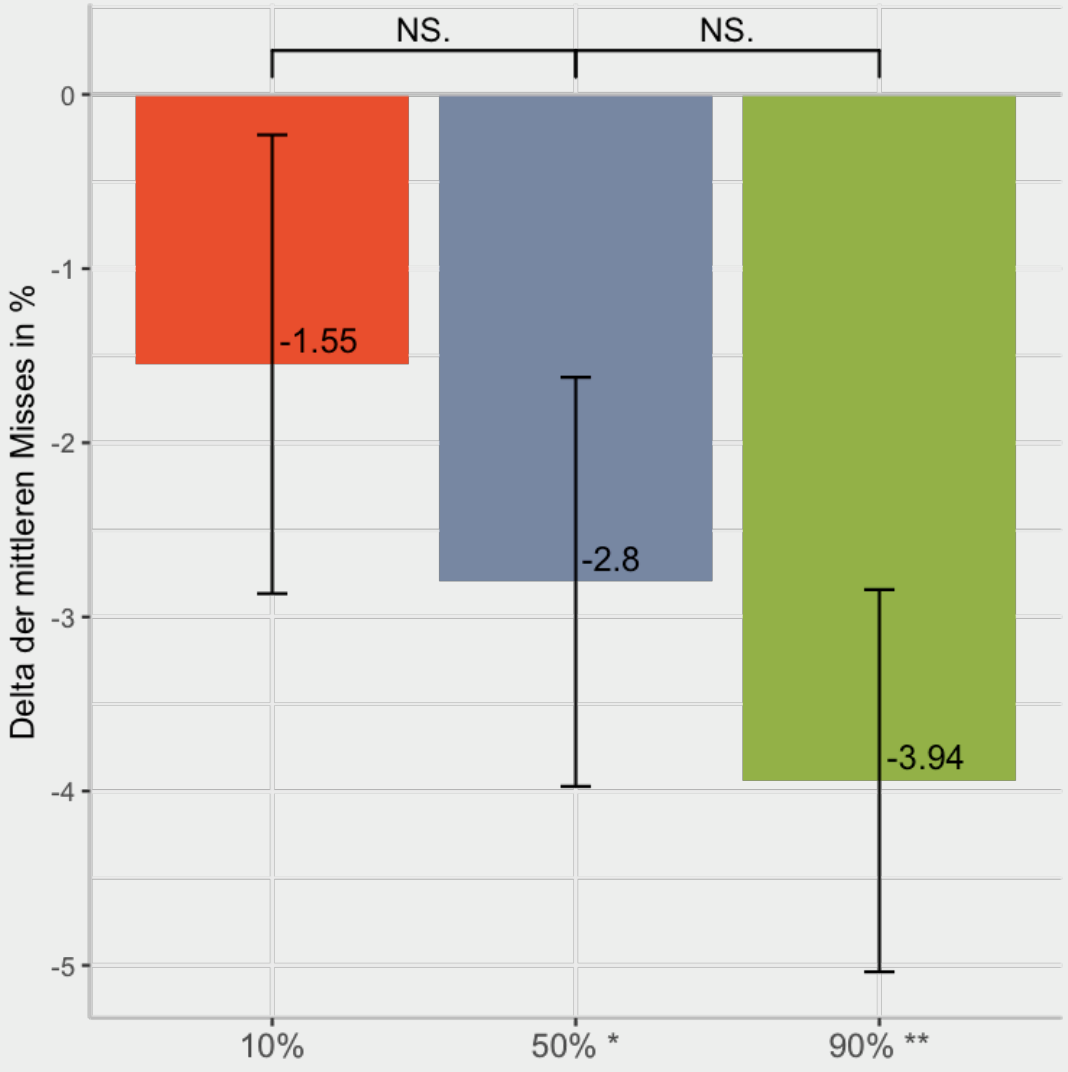


AUSWERTUNG: HYPOTHESEN 14 & 15 (ÜBERSEHENE AS.)

Hypothese * bzgl. der übersehenen AS.	Signifikant? α = 0,025
10% < 50%*	✗
50% < 90%*	✗

- Typ-3 ANOVA (Stufe des Assistenzsystems × Block)
- kein signifikanter Haupteffekt der Stufe des Assistenzsystems
- Haupteffekt des Blocks
- kein signifikanter Interaktionseffekt

- T-Test: 10% vs. 50% und 50% vs. 90%
- ➔ Kein sign. Unterschied zur jeweils benachbarten Stufe



HYPOTHESEN

	Richtigkeit	Tempo	Übersehene Annotationsstellen
10% richtige Assistenz	mehr richtig als ohne	schneller als ohne	weniger übersehen als ohne
50% richtige Assistenz	mehr richtig als ohne	schneller als ohne	weniger übersehen als ohne
90% richtige Assistenz	mehr richtig als ohne	schneller als ohne	weniger übersehen als ohne
10% < 50%	50% richtige Assistenz macht noch mehr richtig	50% richtige Assistenz noch schneller	50% richtige Assistenz noch weniger übersehen
50% < 90%	90% richtige Assistenz macht noch mehr richtig	90% richtige Assistenz noch schneller	90% richtige Assistenz noch weniger übersehen

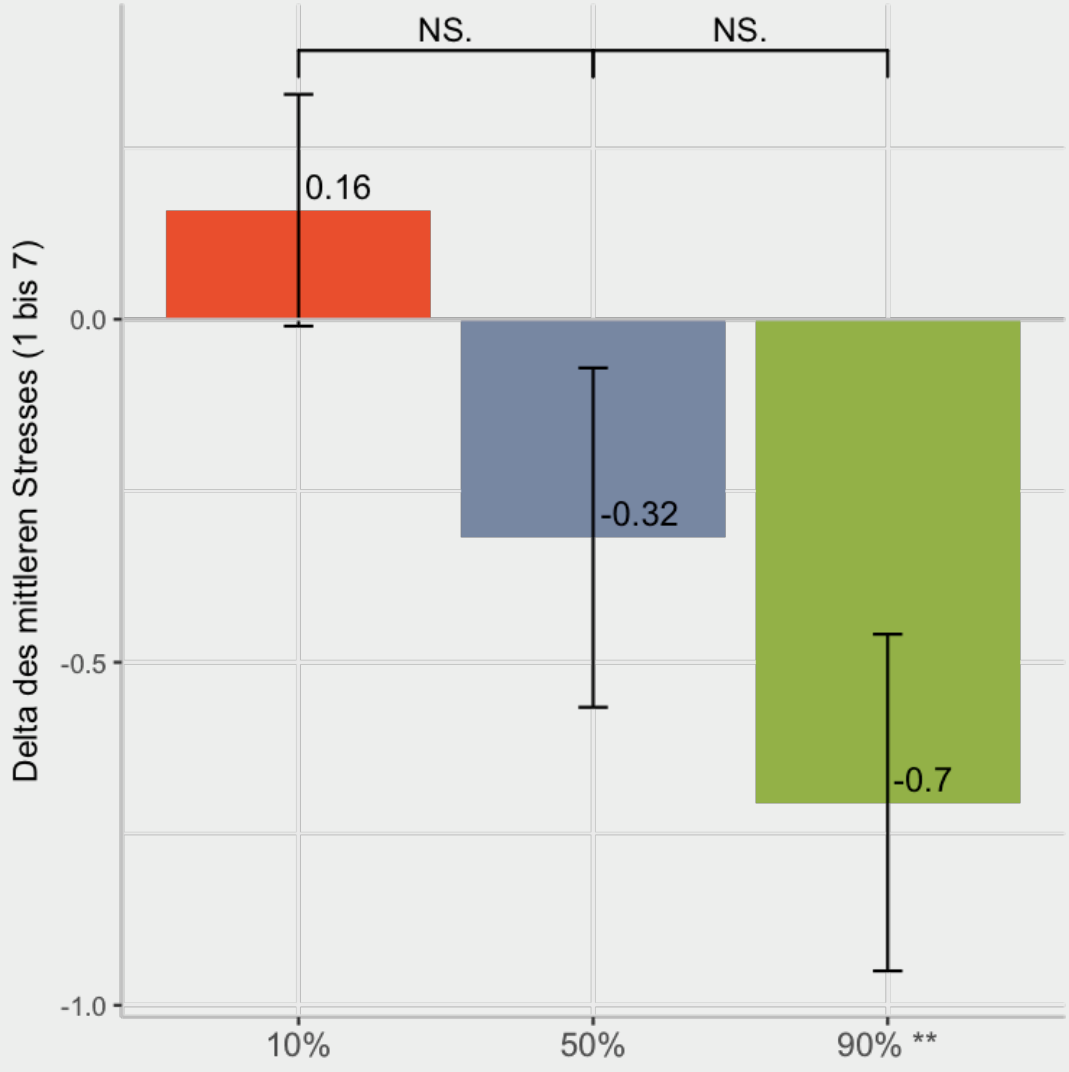
HYPOTHESEN

	Richtigkeit	Tempo	Übersehene Annotationsstellen
10% richtige Assistenz	✗	✗	✗
50% richtige Assistenz	✓	✓	✓
90% richtige Assistenz	✓	✓	✓
10% < 50%	✗	✗	✗
50% < 90%	✗	✗	✗

AUSWERTUNG: PERSÖNLICHE EMPFINDUNGEN

"Wie beansprucht fühlst du dich?"

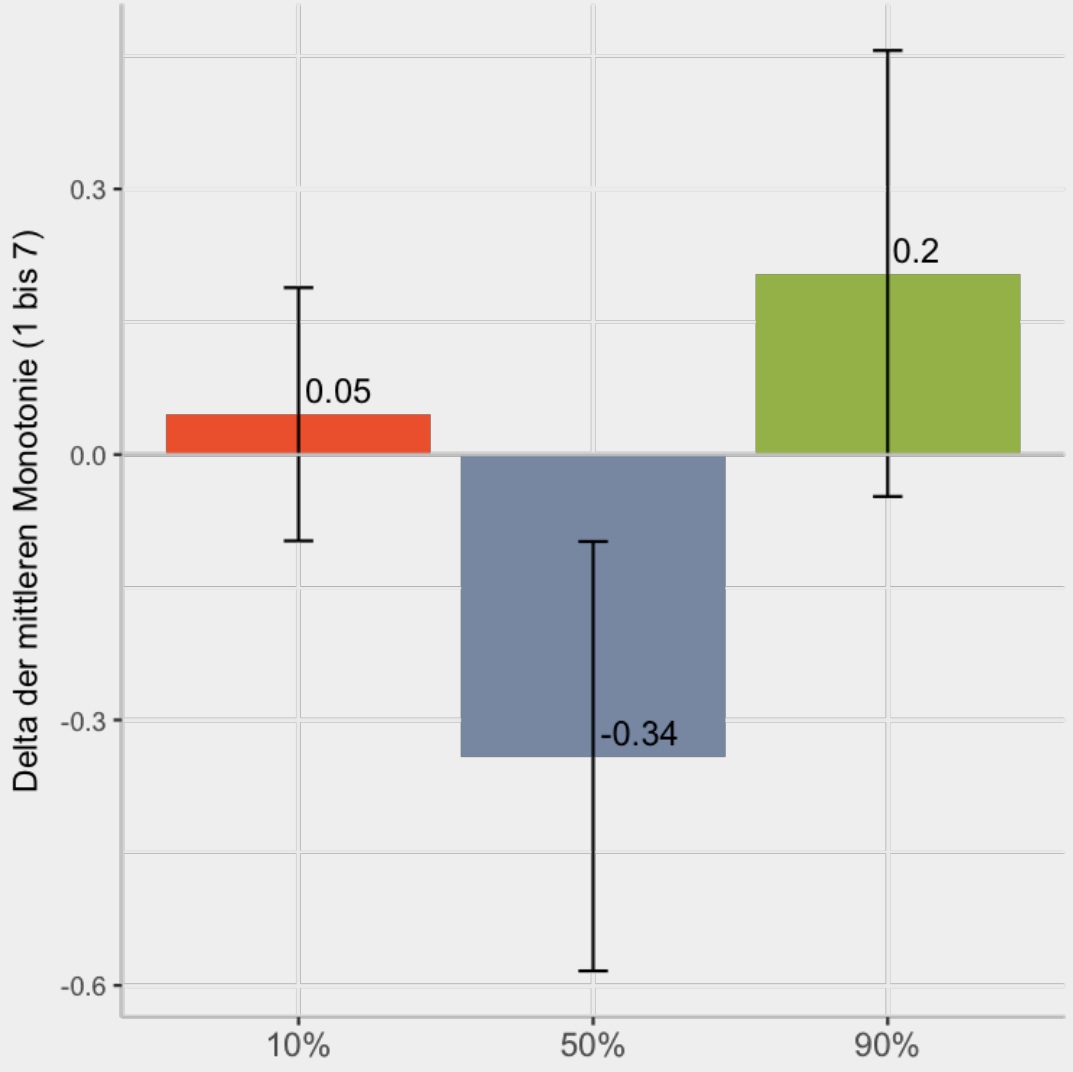
- Baseline: 4,19 auf einer Skala von 1 bis 7
- 3x One Sample T-Test
- ANOVA und Post Hoc Test
- α = 0,05 (Post Hoc: 0,0167)
- ➔ Die Assistenz in der Stufe 90% führt zu signifikant weniger Beanspruchung
- ➔ Kein sign. Unterschied zur jeweils benachbarten Stufe
- r = -0.33



AUSWERTUNG: PERSÖNLICHE EMPFINDUNGEN

"Wie monoton empfandst du die Annotation des vergangenen Blocks?"

- Baseline: 3,67 auf einer Skala von 1 bis 7
- 3x One Sample T-Test
- ANOVA
- α = 0,05
- ➔ Die Assistenz wirkt in keiner Stufe signifikant auf die empfundene Monotonie.
- ➔ Kein sign. Unterschied zur jeweils benachbarten Stufe
- r = 0.06



KONSEQUENZEN

Diskussion, Kritik und
weiterführende Fragestellungen

KONSEQUENZEN

- 10% richtiges Assistenzsystem
  - “Verschlimmbessert“
- 90% richtiges Assistenzsystem
  - unrealistisch
- 50% richtiges Assistenzsystem
  - Technisch realistisch

- ➔ Keine Assistenz wenn Antwortqualität nicht sichergestellt werden kann
  - Nicht von Anfang an
- Genaue Grenze zwischen positivem und negativem Einfluss offen

Weiterführende Fragestellungen

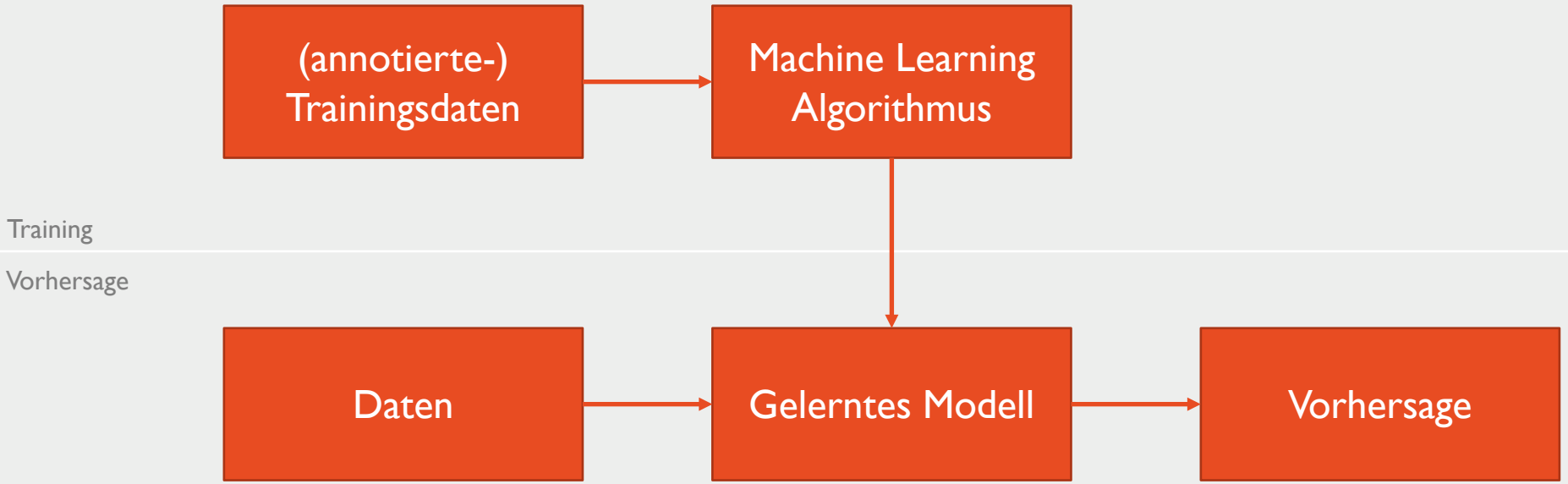
- Linearer Zusammenhang?
  - Ab wann “lohnt“ die Assistenz?
- Unterschiedliche Korrekturleistungen?
  - Wie schwierig ist es unterschiedliche Kategorien von Fehlern zu korrigieren?

TELLERRAND

Wie funktioniert dieses Machine Learning eigentlich?

MACHINE LEARNING

"Ein Teilbereich der künstlichen Intelligenz der Computern die Möglichkeit gibt, durch lernen eine Aufgabe zu lösen ohne speziell darauf programmiert worden zu sein." – Lukas Masuch, SAP



➔ Aus bekannten Daten lernen um anschließend Vorhersagen über neuer Daten zu treffen.