

Bachelorarbeit

Disambiguierungsstrategien in Dialogsystemen

Lena Enzweiler

Universität des Saarlandes

20. Januar 2015

Dialogsystem in automobilen Anwendungen

Effiziente Dialogsysteme im Auto müssen folgende Punkte erfüllen



- Ablenkung während der Fahrt vermeiden
- alle Informationen kurz und verständlich übermitteln
- einfache und intuitive Bedienung garantieren

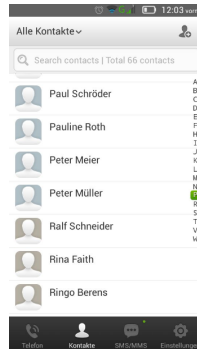
→ Sprachäußerungen müssen durchdacht formuliert werden

Fokus der Studie

- "Rufe Peter an!"
- System muss über Peter Meier und Peter Müller disambiguieren

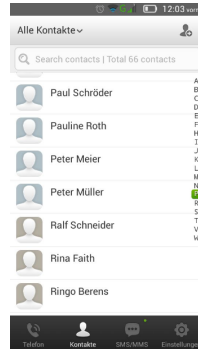
Fokus der Studie

- "Rufe Peter an!"
- System muss über Peter Meier und Peter Müller disambiguieren
- unterschiedliche Disambiguierungsstrategien anwendbar



Fokus der Studie

- "Rufe Peter an!"
- System muss über Peter Meier und Peter Müller disambiguieren
- unterschiedliche Disambiguierungsstrategien anwendbar

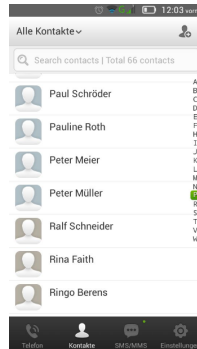


Fokus

→ 3 Disambiguierungsstrategien untersucht

Fokus der Studie

- "Rufe Peter an!"
- System muss über Peter Meier und Peter Müller disambiguieren
- unterschiedliche Disambiguierungsstrategien anwendbar



Fokus

→ 3 Disambiguierungsstrategien untersucht

1. Disambiguierungsstrategie: Aggregierte Auswahl ohne Pause

- alle möglichen Interpretationen in einer Sprachausgabe
- keine Pause zwischen Interpretationen
- Warten auf Auswahl des Benutzers

| Akteur | Sprachausgabe |
|----------|---|
| Benutzer | Rufe Peter an! |
| System | Meinst du Peter Müller oder Peter Meier? |
| Benutzer | Peter Müller. |
| System | Ok, ich werde Peter Müller jetzt anrufen. |

2. Disambiguierungsstrategie: Aggregierte Auswahl mit Pause

- alle möglichen Interpretationen in einer Sprachausgabe
- Pause und Nummerierung zwischen Interpretationen
- Warten auf Auswahl des Benutzers

| Akteur | Sprachausgabe |
|----------|---|
| Benutzer | Rufe Peter an! |
| System | Meinst du [Pause] 1. Peter Müller [Pause] oder 2. Peter Meier? |
| Benutzer | Erstens |
| System | Ok, ich werde Peter Müller jetzt anrufen. |

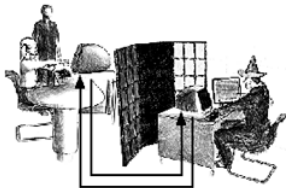
3. Disambiguierungsstrategie: Sequentielle Auswahl

- alle möglichen Interpretationen in einer separaten Sprachausgabe
- Warten auf Zustimmung/Ablehnung des Benutzer

| Akteur | Sprachausgabe |
|----------|--|
| Benutzer | Rufe Peter an! |
| System | Meinst du Peter Müller? |
| Benutzer | Nein. |
| System | Meinst du Peter Meier? |
| Benutzer | Ja. |
| System | Ok, ich werde Peter Meier jetzt anrufen. |

Wizard-of-Oz

- Die Existenz eines funktionierenden Systems wird vorgetäuscht



- Versuchsperson wird der Eindruck verliehen, sie würde mit einem echten Dialogsystem interagieren
- echtes Dialogsystem durch Versuchsleiter simuliert
- Control Panel entwickelt, mit welchem Sprachausgaben ausgegeben werden können

Control Panel

Abbildung: Control Panel

The screenshot shows a window titled 'controlpanel' with a standard Windows-style title bar (minimize, maximize, close buttons). Below the title bar is a tabbed interface with tabs labeled 'commons', 'Anke', 'Peter', 'Fritz', and 'Kim'. The 'Anke' tab is currently selected. On the left side of the main content area, there is a vertical sidebar with a 'Name' label and a 'MobileLandline' label. The main content area contains five numbered steps, each with a question in a text box:

- 1: Soll ich Anke auf der Mobilnummer oder auf der Festnetznummer anrufen?
- 2: Soll ich Anke 1. auf der Mobilnummer oder 2. auf der Festnetznummer anrufen?
- 3A: Soll ich Anke auf der Mobilnummer anrufen?
- 3B: Soll ich Anke auf der Festnetznummer anrufen?
- 3C: Anke hat sonst keine Nummern. Willst du Anke trotzdem anrufen?

At the bottom of the window, there are several buttons for user interaction:

- A 'Success' button followed by a text input field containing 'Ok, ich werde Anke jetzt anrufen'.
- A 'fail' button followed by a text input field containing 'Anke ist leider nicht erreichbar'.
- A 'Cancel' button.
- A 'RequestRepeat' button.
- An 'Okay' button.
- A 'SlotbySlot' button.
- A red 'STOP' button.

Testszenario

- Versuchspersonen sollen vorgegebenen Kontakt anrufen.
- Personenprofil zeigt Informationen über Kontakt
- unspezifische Spracheingabe:
→ Disambiguierung
- pro Anruf unterschiedliche Disambiguierungsstrategie

Anke Schumacher



Mobilnummer

privat

geschäftl.

Festnetznummer

privat

geschäftl.

 Mainzerstr. 23, 66121, Saarbrücken

 A.Schumacher86@gmx.de

Versuchsaufbau

- Versuchspersonen fahren ein Rennspiel. → Fahrsimulation
- Rennspiel: Need for Speed: Shift
- Rennspiel wird mit Lenkrad inklusive Gas- und Bremspedal gespielt → realitätsgetreues Gefühl
- Es wird im Einzelrennen mit jeweils 5 Gegnern gespielt
- Versuchspersonen sollen möglichst hohe Platzierung erreichen
→ Anstrengung und Konzentration soll hohe kognitive Belastung verursachen

Versuchsaufbau - Rennspiel

Abbildung: Need for Speed - Shift



Versuchsaufbau - Überblick

| Vorrunde | 1. Runde | 2. Runde | 3. Runde | 4. Runde |
|-----------|-------------------------|--------------------------|--------------------------|-----------|
| Rennspiel | Rennspiel Anruf Anke | Rennspiel Anruf Peter | Rennspiel Anruf Fritz | Anruf Kim |

- Vorrunde zum Einspielen
- Runde 1-3: Rennspiel mit paralleler Systeminteraktion
→ hohe kognitive Belastung
- Runde 4: nur Systeminteraktion
→ geringe kognitive Belastung

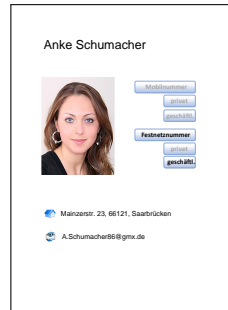
Versuchsdesign

| Aufteilung | Strecke 1 | Strecke 2 | Strecke 3 |
|------------|---------------|---------------|---------------|
| 1. Gruppe | Strategie A | Strategie B | Strategie C |
| 2. Gruppe | Strategie B | Strategie C | Strategie A |
| 3. Gruppe | Strategie C | Strategie A | Strategie B |
| 4. Gruppe | keine Strecke | keine Strecke | keine Strecke |

- 3 verschiedene Strecken, um Lerneffekt auszuschließen
- Die Strecken werden in gleicher Reihenfolge gefahren
- jede Strecke mit unterschiedlicher Disambiguierungsstrategie
- um Zeiten besser zu vergleichen:
 - Disambiguierungsstrategien werden auf Strecken verteilt
 - Versuchspersonen werden in Gruppen (1-3) aufgeteilt
- Gruppe 4 führt das Testszenario mit zufälliger Strategie aus.

Testszenario

- Testperson soll 4 Anrufe aufbauen
- Pro Anruf: Disambiguierung über zwei Merkmale
- Disambiguierung erfolgt mit 2 Alternativen
- Personenprofil zeigt zu füllende Disambiguierungsmerkmale



Beispiel: Disambiguierung über Namen

Benutzer: Rufe Anke an

System: Meinst du Anke Meier oder Anke Schuhmacher?

Benutzer: Schuhmacher

Versuchspersonen

- 12 deutsche Muttersprachler
- 58% 18-29 Jahre, 17% 30-41 Jahre, 25% 42-53 Jahre
- 75% keine bzw. wenig Erfahrung mit Dialogsystemen
- 83% spielen selten Rennspiele
- 58% fiel Einführungsrunde schwer

Auswertung

Folgende Punkte werden ausgewertet

- Zeiten werden gemessen
 - Rennzeiten
 - Dialogzeiten
- Fragebögen ausgewertet
 - Nasa-TLX
 - Strategien
- Task Completion
- Dialogverhalten

Gemessene Zeiten - Rennzeiten

Rennzeiten

Beeinflusst eine Disambiguierungsstrategie das Rennverhalten?

| Rennzeiten | Strategie 1 | Strategie 2 | Strategie 3 |
|--------------|-------------|-------------|-------------|
| Durchschnitt | 71,58 sek | 75,71 sek | 75,92 sek |

Zeiten statistisch nicht relevant und daher nicht aussagekräftig.

Gemessene Zeiten - Dialogzeiten

Dialogzeiten

Welche Strategie ermöglicht den kürzesten Dialog?

Nur die Zeiten von korrekt durchgeführten Dialogen bewertet.

| Dialogzeiten | Strategie 1 | Strategie 2 | Strategie 3 |
|----------------|-------------|-------------|-------------|
| mit Rennspiel | 15,2 sek | 20,5 sek | 20,8 sek |
| ohne Rennspiel | 14,9 sek | 18,8 sek | 17,6 sek |

⇒ Strategie 1 ermöglicht den kürzesten Dialog.

Gemessene Zeiten - Dialogzeiten

Dialogzeiten

Gibt es Unterschiede in den Dialogzeiten zwischen kognitiv hoch belasteten und kognitiv wenig belasteten Versuchspersonen?

| Dialogzeiten | Strategie 1 | Strategie 2 | Strategie 3 |
|----------------|-------------|-------------|-------------|
| mit Rennspiel | 15,2 sek | 20,5 sek | 20,8 sek |
| ohne Rennspiel | 14,9 sek | 18,8 sek | 17,6 sek |

- kürzere Dialogzeiten ohne Rennspiel erreicht

⇒ bessere Reaktionszeit unter geringer Belastung

Fragebogen - Nasa-TLX

Nasa-TLX

- ❶ Bei welcher Strategie wurde höhere Belastung empfunden?
- ❷ Gibt es Unterschiede in der Belastung zwischen den Runden mit und ohne Rennspiel?

- geistige Anforderung
 - Strategie 1: geringe geistige Anforderung
 - Strategie 3: höchste geistige Anforderung
 - Runde ohne Rennspiel weniger anfordernd
- Anstrengung
 - Strategie 1: geringe Anstrengung
 - Strategie 3: höchste Anstrengung
 - Runde ohne Rennspiel weniger anstrengend

⇒ Strategie 1 am wenigsten belastend gewertet

Fragebogen - Strategien

Strategien

Wie werden die Strategien von den Versuchspersonen bewertet?

- Strategie 1 lenkte am wenigsten ab
- Dialog aus Strategie 1 gefiel am besten, Strategie 3 am schlechtesten
- Strategie 1 wurde von 75% als beste Strategie gewählt (17% Strategie 2, 8% Strategie 1)

→ Strategie 1 insgesamt am besten bewertet

- der Dialog fiel im Durchschnitt ohne Rennspiel einfacher

Task Completion

Task Completion (TC)

Welche Strategie ist am erfolgversprechendsten?

- Die Task Completion wird für jeden Dialog wie folgt berechnet:
 - 0 Punkte, wenn kein Slot richtig gefüllt wird
 - 1 Punkt, wenn ein Slot richtig gefüllt wird
 - 2 Punkte, wenn alle Slots richtig gefüllt werden
- für jede Strategie wird die durchschnittliche Task Completion bewertet

Task Completion

| Strategien | Runde 1-4 | Runde 1-3 | Runde 4 |
|---------------------|-----------|-----------|---------|
| 1. Strategie | 1,75 | 1,92 | 1,50 |
| 2. Strategie | 1,94 | 1,92 | 2,00 |
| 3. Strategie | 1,63 | 1,50 | 2,00 |
| - alle Strategien - | - | - | - |
| | | 1,78 | 1,83 |

- Strategie 2 am erfolgreichsten
- Strategie 3 am unerfolgreichsten
- Runde ohne Rennspiel erfolgreicher als Runden mit Rennspiel

→ Unterschied gering: 0.05

Dialogverhalten

Dialogverhalten

Gibt es Unterschiede im Dialogverhalten bei unterschiedlicher Belastung

- Antwort aus Runde 1-3 mit Antwort aus Runde 4 verglichen
- 11 von 12 Personen wiesen gleiches Verhalten auf

⇒ kein unterschiedliches Dialogverhalten

Auswertung - Zusammenfassung

- **kürzeste Dialogzeit:** Strategie 1
- Ergebnis **Nasa-TLX** Fragebogen:
 - Strategie 1 am unbelastetsten
 - Runde ohne Rennspiel weniger belastend als Runde mit
- Ergebnis **Strategien** Fragebogen:
 - Strategie 1 am positivsten bewertet
 - Dialog fiel im Durchschnitt ohne Rennspiel einfacher
- **Task Completion**
 - Strategie 2 > Strategie 1 > Strategie 3 (geringer Unterschied)
 - Dialog ohne Rennspiel erfolgreicher als Dialog mit Rennspiel
- **Dialogverhalten:** kein unterschiedliches Dialogverhalten bei unterschiedlicher Belastung

⇒ **Strategie 1** am Effizientesten und Beliebtesten

Versuch 2

Versuch 1 zeigte eindeutiges Ergebnis bei Disambiguierung über 2 Alternativen

Fragestellung

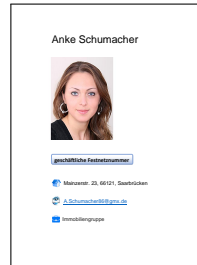
Gleiches Ergebnis bei Disambiguierung über mehr Alternativen?

→ Zweiter Versuch eingeleitet:

- gleiche Versuchsdurchführung
- Unterschied zu Versuch 1: längere Disambiguierung

Testszenario

- Testperson soll 4 Anrufe aufbauen
- Pro Anruf: Disambiguierung über zwei Merkmalen
- Disambiguierung erfolgt mit 6 Alternativen
- Personenprofil zeigt zu füllende Disambiguierungsmerkmale



Beispiel: Disambiguierung über Namen

Benutzer: Rufe Anke an

System: Meinst du Anke Bies, Anke Elb, [...] oder Anke Weiler?

Benutzer: Schuhmacher

Versuchspersonen

- 12 deutsche Muttersprachler
 - 42% 18-29 Jahre, 25% 30-41 Jahre, 33% 42-53 Jahre
 - 75% keine bzw. wenig Erfahrung mit Dialogsystemen
 - 83% spielen selten Rennspiele
 - 58% fiel Einführungsrunde schwer
- Zufällig gleiche Erfahrungswerte wie in Versuch 1:
- Unterschiedliche Resultate der Versuche nicht durch unterschiedliche Erfahrung zu erklären

Auswertung

Folgende Punkte werden ausgewertet

- Zeiten werden gemessen
 - Rennzeiten
 - Dialogzeiten
- Fragebögen ausgewertet
 - Nasa-TLX
 - Strategien
- Task Completion
- Dialogverhalten

Gemessene Zeiten - Rennzeiten

Rennzeiten

Beeinflusst eine Disambiguierungsstrategie das Rennverhalten?

| Rennzeiten | Strategie 1 | Strategie 2 | Strategie 3 |
|--------------|-------------|-------------|-------------|
| Durchschnitt | 76,83 sek | 77,47 sek | 76,73 sek |

Zeiten statistisch nicht relevant und daher nicht aussagekräftig.

Gemessene Zeiten - Dialogzeiten

Dialogzeiten

Welche Strategie ermöglicht den kürzesten Dialog?

Nur die Zeiten von korrekt durchgeführten Dialogen bewertet.

| Dialogzeiten | Strategie 1 | Strategie 2 | Strategie 3 |
|----------------|-------------|-------------|-------------|
| mit Rennspiel | 29,6 sek | 38,5 sek | 34,3 sek |
| ohne Rennspiel | 24,1 sek | 34,4 sek | 30,4 sek |

→ Strategie 1 ermöglicht den kürzesten Dialog.

Gemessene Zeiten - Dialogzeiten

Dialogzeiten

Gibt es Unterschiede in den Dialogzeiten zwischen kognitiv hoch belasteten und kognitiv wenig belasteten Versuchspersonen?

| Dialogzeiten | Strategie 1 | Strategie 2 | Strategie 3 |
|----------------|-------------|-------------|-------------|
| mit Rennspiel | 29,6 sek | 38,5 sek | 34,3 sek |
| ohne Rennspiel | 24,1 sek | 34,4 sek | 30,4 sek |

- kürzere Dialogzeiten ohne Rennspiel erreicht

→ bessere Reaktionszeit bei Dialoginteraktion ohne Rennspiel

Fragebogen - Nasa-TLX

Nasa-TLX

- ❶ Bei welcher Strategie wurde eine höhere Belastung empfunden?
- ❷ Gibt es Unterschiede in der Belastung zwischen den Runden mit und ohne Rennspiel?

- geistige Anforderung
 - Strategie 3: höchste geistige Anforderung
 - Runde ohne Rennspiel weniger anfordernd
- Anstrengung
 - Strategie 3: geringe Anstrengung
 - Strategie 2: höchste Anstrengung
 - Runde ohne Rennspiel weniger anstrengend

→ Unterschiedliche empfundene Belastung mit und ohne Rennspiel
→ keine Aussage über belastendste Strategie treffbar.

Fragebogen - Strategien

Strategien

Wie werden die Strategien von den Versuchspersonen bewertet?

- Strategie 2 lenkte am wenigsten ab, Strategie 1 am meisten
- Dialog aus Strategie 3 gefiel am besten
- Strategie 3 wurde von 50% als beste Strategie gewählt (17% Strategie 1, 33% Strategie 2)

→ Strategie 3 am beliebtesten bewertet
→ keine Strategie eindeutig am besten bewertet

- der Dialog fiel im Durchschnitt ohne Rennspiel einfacher

Task Completion

Task Completion (TC)

Welche Strategie ist am erfolgversprechendsten?

- Die Task Completion wird für jeden Dialog wie folgt berechnet:
 - 0 Punkte, wenn kein Slot richtig gefüllt wird
 - 1 Punkt, wenn ein Slot richtig gefüllt wird
 - 2 Punkte, wenn alle Slots richtig gefüllt werden
- für jede Strategie wird die durchschnittliche Task Completion bewertet

Task Completion

| Strategien | Runde 1-4 | Runde 1-3 | Runde 4 |
|-----------------|-----------|-----------|---------|
| 1. Strategie | 1,88 | 1,83 | 2,00 |
| 2. Strategie | 1,81 | 1,75 | 2,00 |
| 3. Strategie | 1,56 | 1,42 | 2,00 |
| alle Strategien | | 1,67 | 2,00 |

- Strategie 1 am erfolgreichsten
- Strategie 3 am unerfolgreichsten
- Runde ohne Rennspiel erfolgreicher als Runden mit Rennspiel

Dialogverhalten

Dialogverhalten

Gibt es Unterschiede im Dialogverhalten bei unterschiedlicher Belastung

- Antwort aus Runde 1-3 mit Antwort aus Runde 4 verglichen
- 11 von 12 Personen wiesen gleiches Verhalten auf

⇒ kein unterschiedliches Dialogverhalten

Auswertung - Zusammenfassung

- **kürzeste Dialogzeit:** Strategie 1
- Ergebnis **Nasa-TLX** Fragebogen:
 - kein eindeutiges Ergebnis über belastendste Strategie
 - Runde ohne Rennspiel weniger belastend als Runde mit
- Ergebnis **Strategien** Fragebogen:
 - keine Strategie eindeutig am besten bewertet
 - Strategie 3 am beliebtesten
 - Dialog fiel im Durchschnitt ohne Rennspiel einfacher
- **Task Completion**
 - Strategie 1 > Strategie 2 > Strategie 3 (geringer Unterschied)
 - Dialog ohne Rennspiel erfolgreicher als Dialog ohne Rennspiel
- **Dialogverhalten:** kein unterschiedliches Dialogverhalten bei unterschiedlicher Belastung

⇒ Strategie 1 am Effizientesten

⇒ Strategie 3 am Beliebtesten

Gemessene Zeiten - Rennzeiten

Rennzeiten

Beeinflusst eine Disambiguierungsstrategie das Rennverhalten?

- Versuch 1: Zeiten nicht aussagekräftig
- Versuch 2: Zeiten nicht aussagekräftig

Gemessene Zeiten - Dialogzeiten

Dialogzeiten

Welche Strategie ermöglicht den kürzesten Dialog?

- Versuch 1: Strategie 1
- Versuch 2: Strategie 1

⇒ Strategie 1 ermöglicht in beiden Versuchen den kürzesten Dialog.

Gemessene Zeiten - Dialogzeiten

Dialogzeiten

Gibt es Unterschiede in den Dialogzeiten zwischen kognitiv hoch belasteten und kognitiv wenig belasteten Versuchspersonen?

- Versuch 1: kürzerer Dialog ohne Rennspiel
- Versuch 2: kürzerer Dialog ohne Rennspiel

⇒ kürzerer Dialog bei geringerer Belastung in beiden Versuchen

Fragebogen - Nasa-TLX

Nasa-TLX

- 1 Bei welcher Strategie wurde eine höhere Belastung empfunden?
- 2 Gibt es Unterschiede in der Belastung zwischen den Runden mit und ohne Rennspiel?

- Versuch 1:
 - Strategie 1 am wenigsten belastend gewertet
 - Runde ohne Rennspiel weniger belastend
- Versuch 2:
 - keine Aussage über belastendste Strategie treffbar
 - Runde ohne Rennspiel weniger belastend

→ Unterschiedlich empfundene Belastung mit und ohne Rennspiel
→ Strategien in beiden Versuchen unterschiedlich bewertet

Fragebogen - Strategien

Strategien

Wie werden die Strategien von den Versuchspersonen bewertet?

- Versuch 1:
 - Strategie 1 eindeutig am besten gewertet
- Versuch 2:
 - keine Strategie eindeutig am besten gewertet
 - Strategie 3 jedoch am beliebtesten

→ Strategien in beiden Versuchen unterschiedlich beliebt

Task Completion

Task Completion (TC)

Welche Strategie ist am erfolgversprechendsten?

- Versuch 1:
 - Strategie 2 am erfolgreichsten
 - Runde 4 erfolgreicher als Runden 1-3
- Versuch 2:
 - Strategie 1 am erfolgreichsten
 - Runde 4 erfolgreicher als Runden 1-3

→ Task Completion unterschiedlich in beiden Versuchen

→ Runde mit geringer Belastung am erfolgreichsten

Dialogverhalten

Dialogverhalten

Gibt es Unterschiede im Dialogverhalten bei unterschiedlicher Belastung

- Versuch 1: kein unterschiedliches Dialogverhalten
- Versuch 2: kein unterschiedliches Dialogverhalten

Auswertung - Zusammenfassung

Erkenntnis





Länge der Disambiguierung beeinflusst Strategienbeliebtheit

- Disambiguierung mit wenigen Alternativen:
 - ⇒ Strategie 1 beliebt und effizient
- Disambiguierung mit vielen Alternativen:
 - ⇒ Strategie 1 effizient
 - ⇒ Strategie 3 beliebt

Vielen Dank

Vielen Dank für Ihre Aufmerksamkeit

Quellen

-  Hervé Abdi und Lynne J. Williams: *Newmann-Keuls Test and Tukey Test* (2006).
-  Daniel Jurafsky und James Martin: *Speech and Language Processing* Prentice Hall 2. Auflage (2008).
-  Jessica Villing: *Dialogue behaviour under high cognitive load* Proceedings of SIGDIAL 2009: the 10th Annual Meeting of the Special Interest Group in Discourse and Dialogue, pages 322–325,(2009)
-  P. Tsiakoulis, M. Henderson, B. Thomson, K. Yu, E. Tzirkel, S. Young: *The Effect of Cognitive Load on a Statistical Dialogue System* Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), pages 74–78, Seoul, South Korea, (July 2012).

Statistische Signifikanz

- einfache Varianzanalyse (one-way ANOVA):
Überprüfung ob es Unterschiede in den Ergebnissen für die einzelnen Strategien gibt
- Tukey Test:
Überprüfung, zwischen welchen Strategien es signifikante Unterschiede gibt

$$\frac{M_1 - M_2}{\sqrt{MS_{error}(\frac{1}{n})}}$$

M_1 und M_2 enthalten die Mittelwerte der Strategien.
 MS_{error} beinhaltet den quadratischen Mittelwert.
 n ist die Anzahl der Durchführungen.