Voice User Interface Design für Anfänger

Grundlagen des VUI Designs

World Usability Day 2019
Jana Paulick

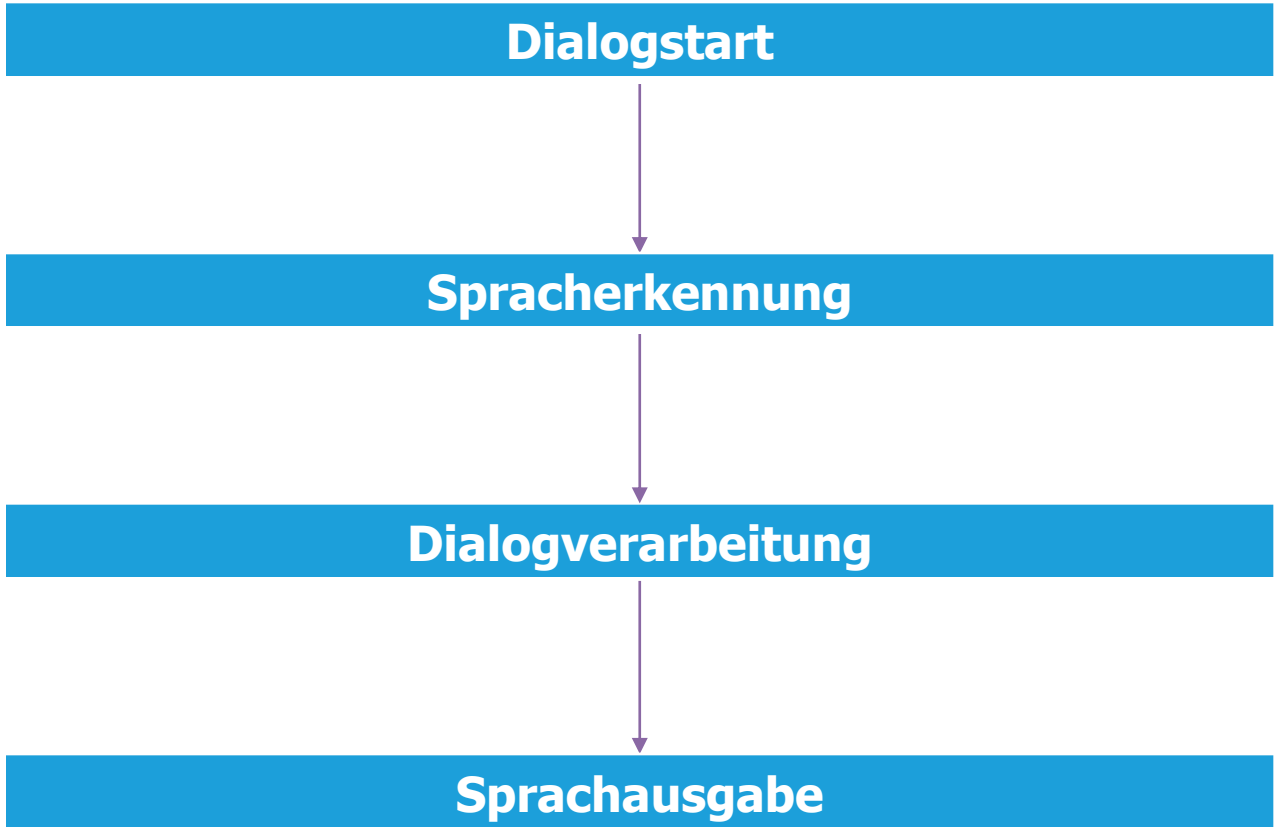
Kurzvorstellung – Jana Paulick

- geb. 1984 in Spremberg
- 2003-2009 Studium Informations- und Medientechnik an der BTU Cottbus
- Master of Science

- seit 2011 beim Spiegel Institut Ingolstadt
  - als Consultant User Experience
  - seit 2019 als Expert VUI Design
  - hauptsächlich für die Audi AG

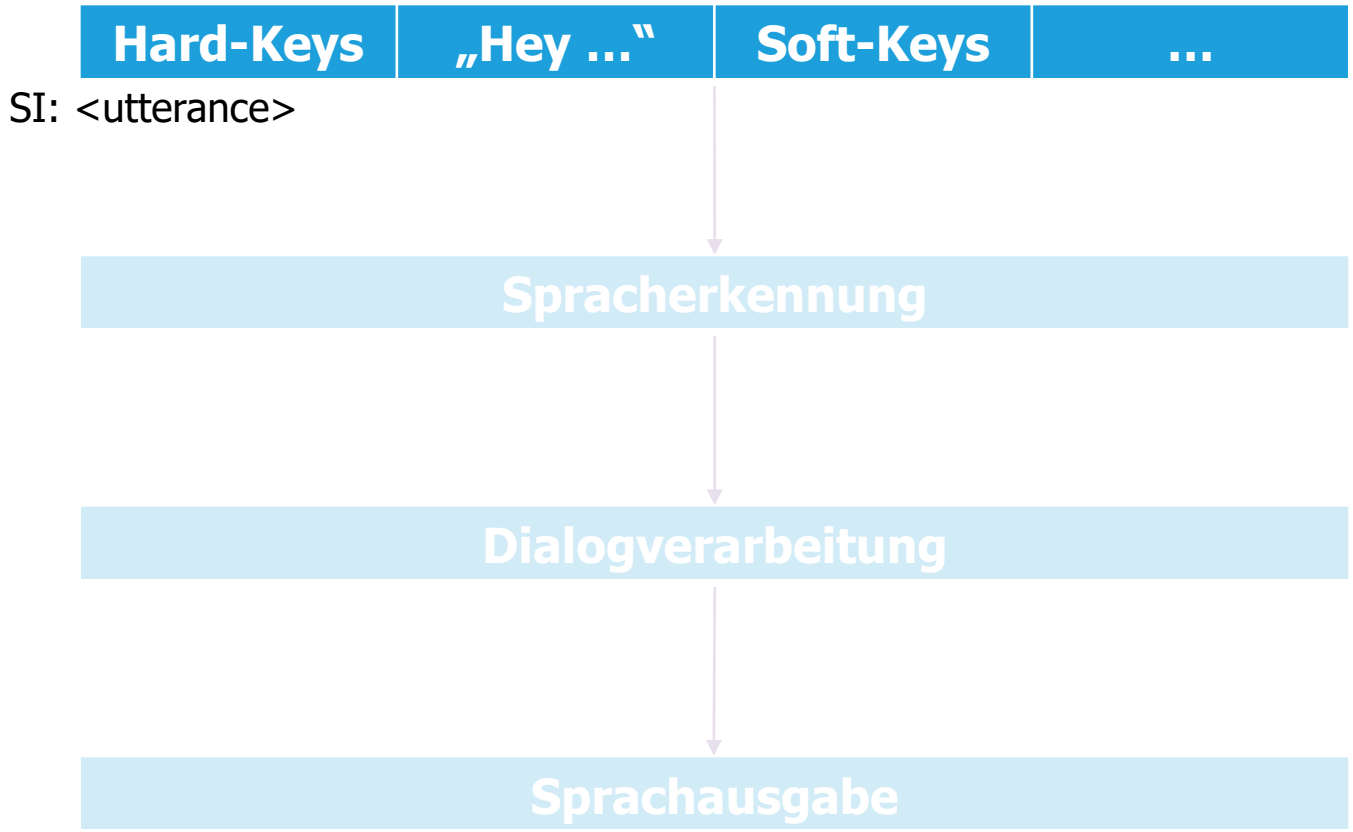
- Schwerpunkte
  - Navigation, Messaging & Online-Dienste
  - fürs Audi MMI (bisher 3 Generationen)

Grundsätzliche Funktionsweise des Sprachdialoges



Dialogstart

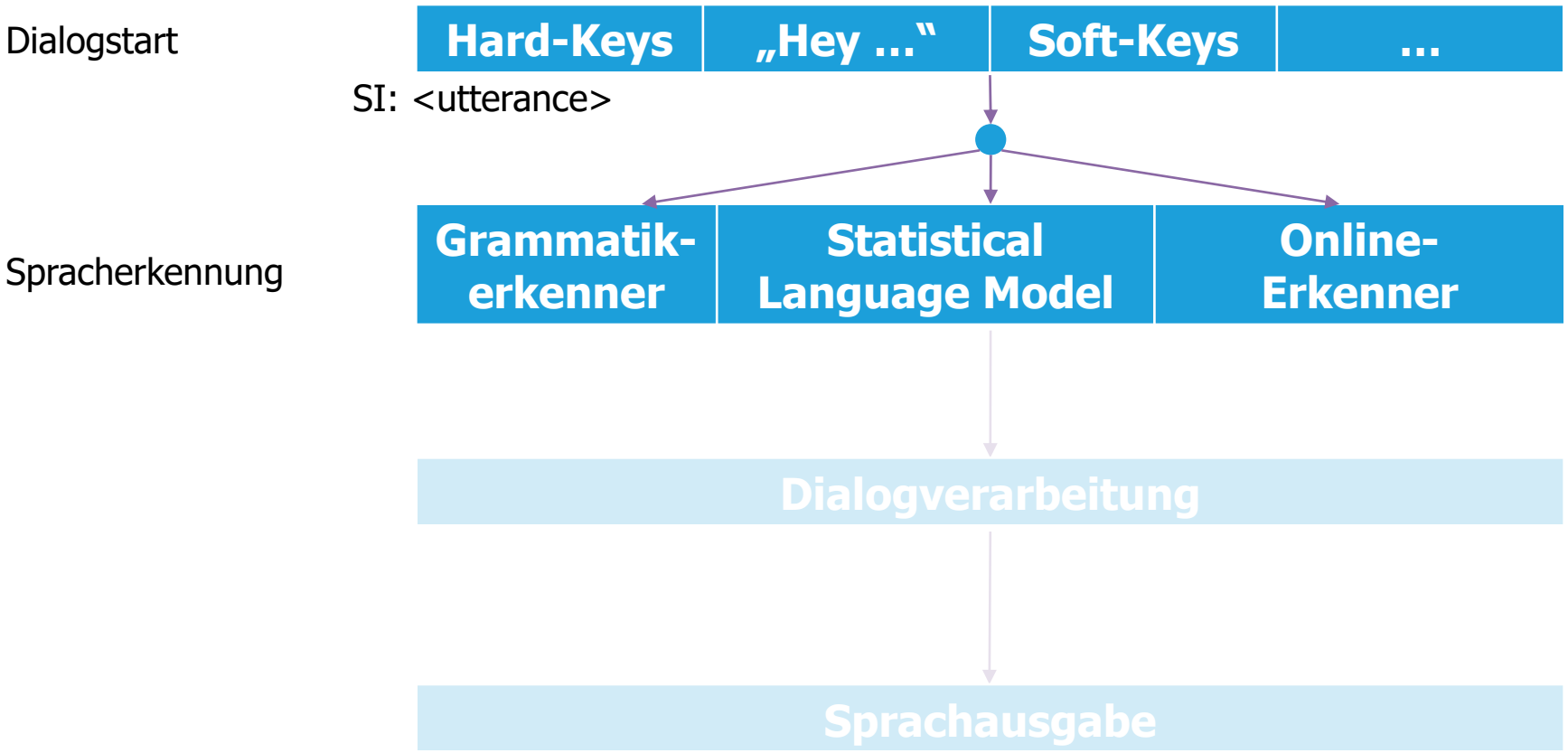
Dialogstart



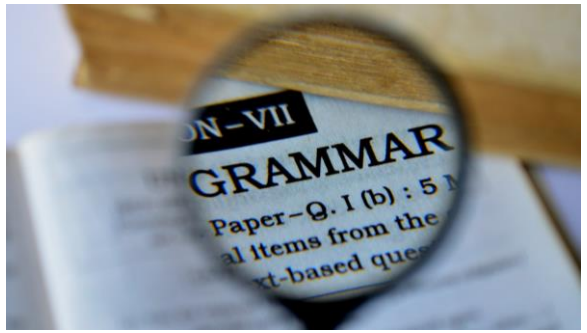
Wie kann man einen Sprachdialog starten?

	Im Auto	Im/am Smartphone
Keyword-Activation ("Hey ...")	x	x
PTT-Taste am Lenkrad	x	-
(Short)-Press auf dedizierte Icons	x	x
(Long)-Press auf Hard-Keys	(-)	x

Spracherkennung



Relevante Spracherkennungstechnologien



Grammatik-Erkenner

- läuft onboard
- Erkennung basiert auf Mustervergleich mittels SRGS-Grammatiken
- klare Vorgabe des Gesagten
- vorhandene Adressbuchnamen können als Guestkontext geladen werden
- Training von (sicherheitsrelevanten) Funktionen/Eigennamen im Fahrzeug kontrollierbarer

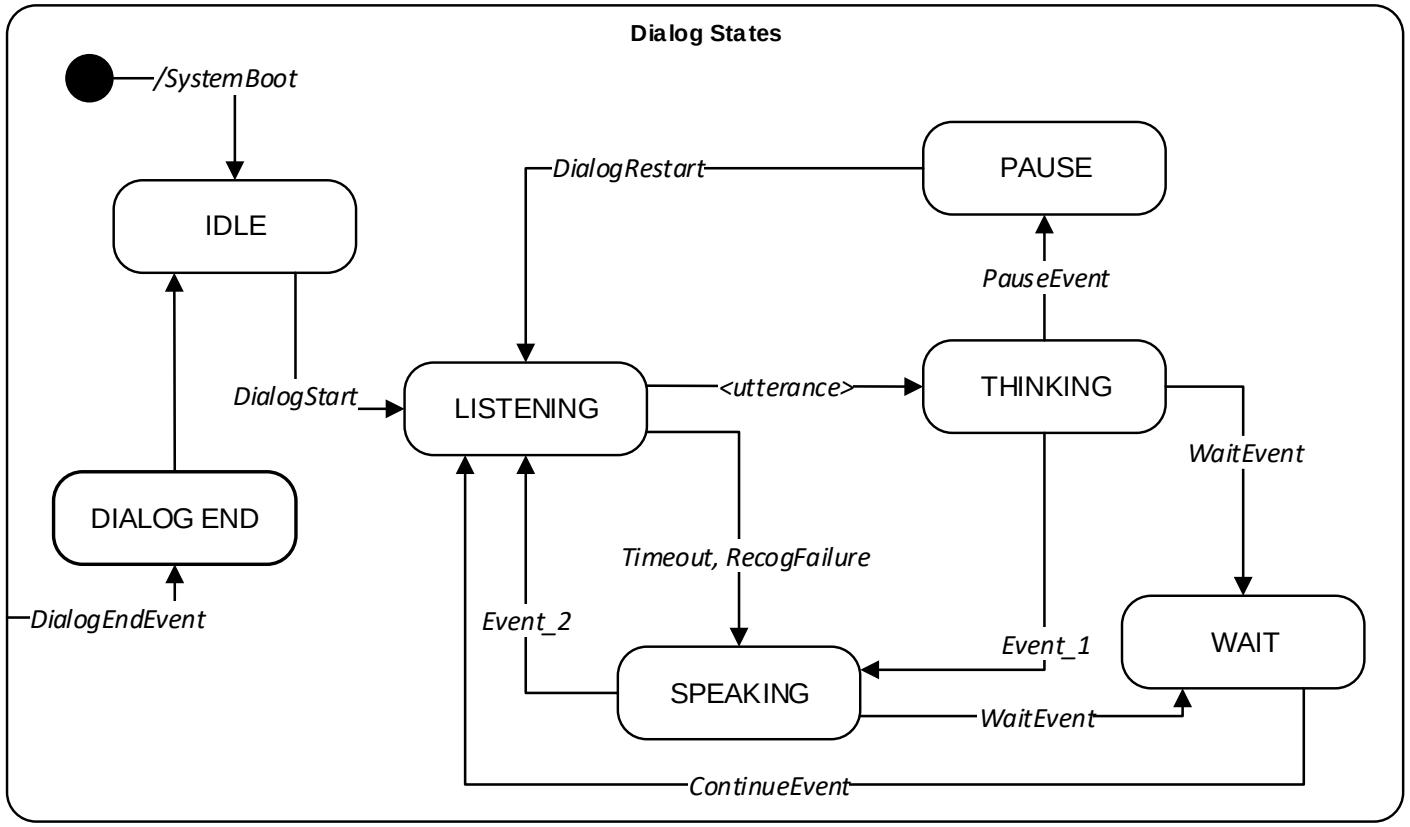
Statistical Language Modell (SLM)

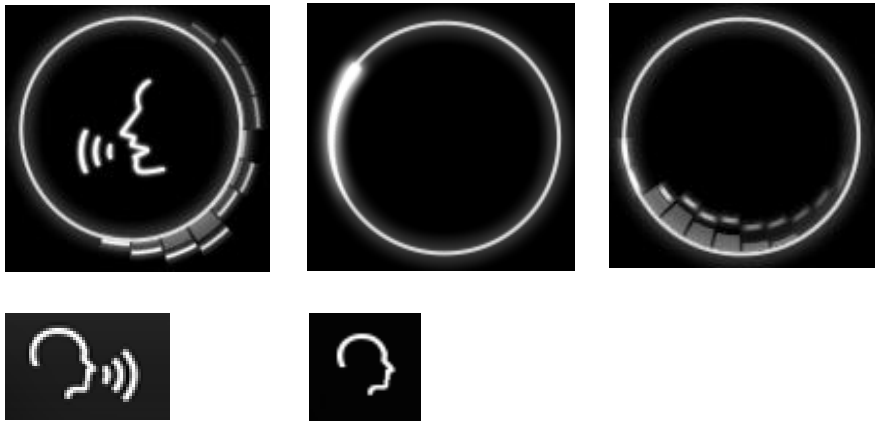
- läuft onboard
- höhere Varianz des Gesagten
- dynamischere Erkennung basierend auf Statistiken
- vorhandene Adressbuchnamen können als Guestkontext geladen werden

Online-Erkenner

- hohe Varianz des Erkennbaren auch bei großen Datenmengen
- Erkennungsraum ist nicht durch Grammatiken eingeschränkt
- Kein Abgleich gegen vorhandene Adressbuchnamen
- Guestkontexte sind nicht gut unterstützt
- häufig Speech-to-Text-Erkennung
- Abgleich zwischen Rückgabewerten der OnlineASR und NavDB notwendig

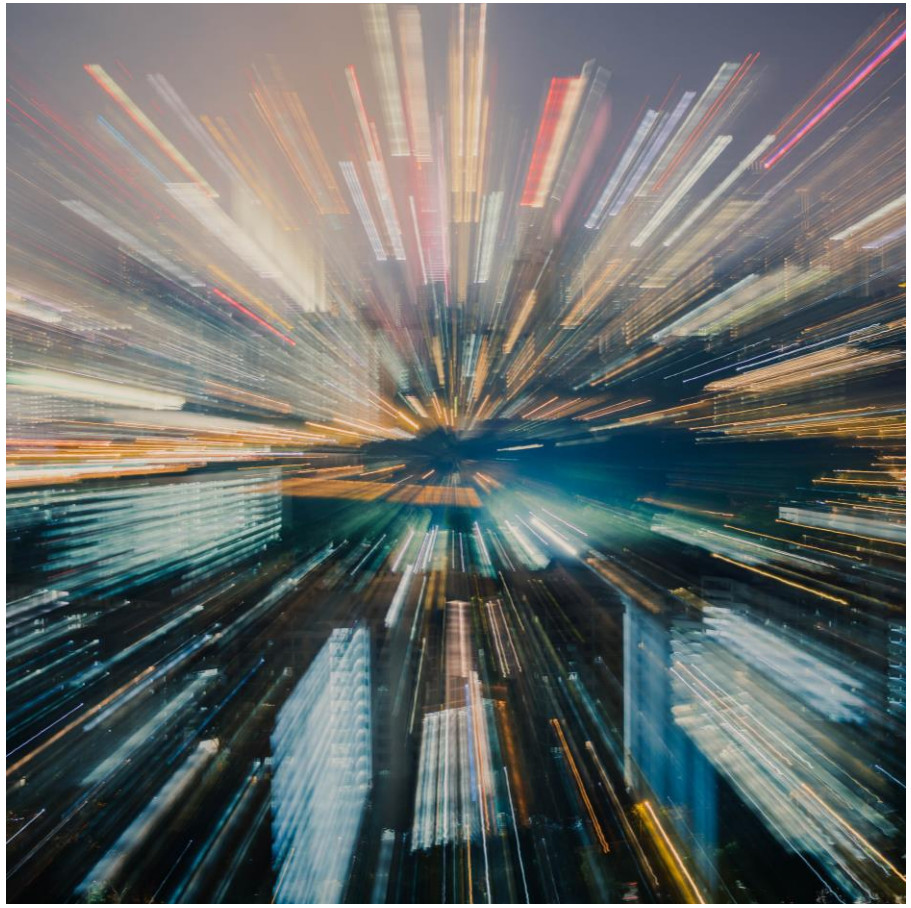
Mögliche Zustände des Dialoges & Spracherkenners





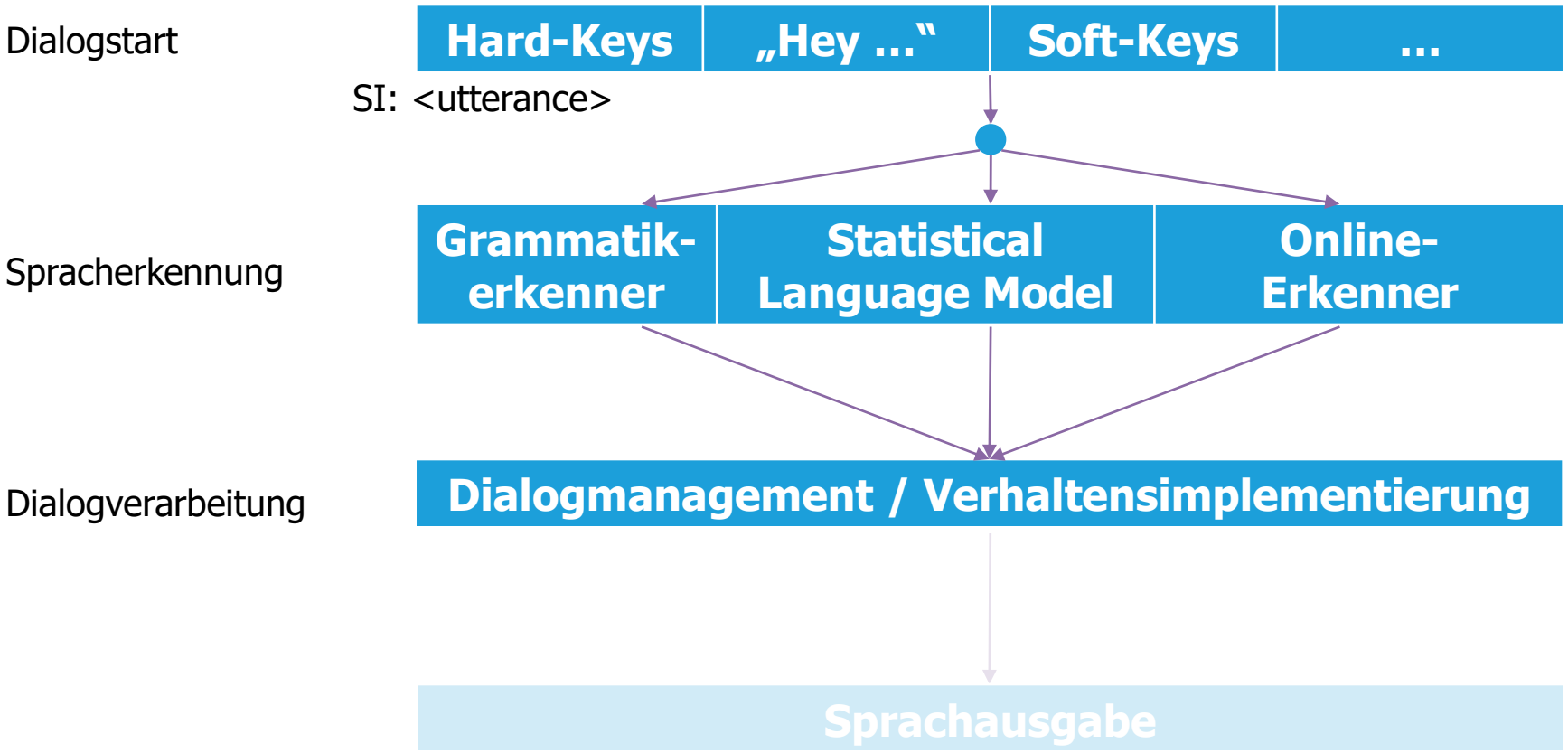
Dynamische Anteile in Nutzeräußerungen

- Stellung von statischen und dynamischen Anteilen
  - en: „call John Doe“ / „call <first name> <last name>“
  - de: „Max Mustermann anrufen / „<Vorname> <Nachname> anrufen“
- Reihenfolge von dynamischen Anteilen
  - Straße, HNR, Stadt
  - Stadt, Straße, HNR
  - HNR, Straße, Stadt
- Nutzung von Füllwörtern
  - SI: „McDonalds **in** Hamburg“
  - SI: „fahre mich nach Hamburg **in die** Edmund-Siemers-Allee“



spiegelInstitut

Dialogverarbeitung



Dialogmanagement / Verhaltensimplementierung

Dialogmanagement / Verhaltensimplementierung

Pause-Verhalten

- automatische Pausenaktivierung:
  - bei eingehenden Telefonanruf
  - beim Scrollen in Listen
- explizite Pausenaktivierung:
  - Druck auf Erkennerstatus - Play-Pause-Toggle
  - SI: Pause

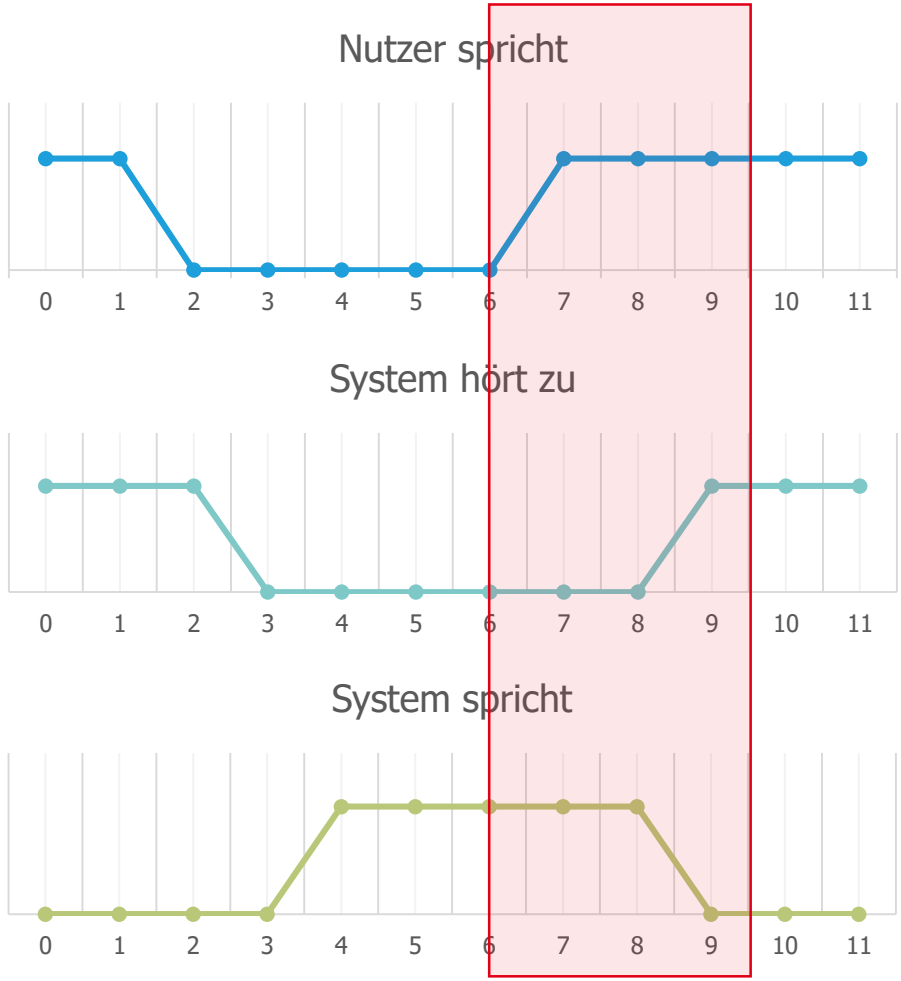
Dialogreaktivierung

- Druck auf Erkennerstatus - Play-Pause-Toggle
- Druck auf PTT*
- Druck auf Listenitem
- Nicht vergessen: Auch eine explizit aufgerufene Pause muss beendet werden!

- PTT: „Push-to-talk" beschreibt die Nutzerinteraktion indem der Nutzer eine Taste oder einen Button drückt und wieder los lässt), um danach zu sprechen.

Voice Bargein*

- Problemstellung:
  - Nutzer warten nicht bis das *Beep* ertönt
  - → sprechen zu früh
- Folge:
  - Erkenner hört nicht, wenn Nutzer spricht
  - → Fehlerkennung
- Mögliche Lösung:
  - PTT-Bargein beendet den *Prompt* vorzeitig
  - →Vorgehen ist vielen Nutzern unbekannt
  - → daher nicht intuitiv
- Alternativ:
  - Voice-Bargein bzw. Nutzer redet, wann es ihm beliebt
  - → *Beep* muss rausgefiltert werden
  - → nicht trivial



* Voice-Bargein: „sprachliches Einmischen" (engl. „barge in" = einmischen).
Der Nutzer kann sich über eine sprachliche Interaktion in die Promptausgabe des Systems „einmischen" und diese Unterbrechen.

Dialogmanagement / Verhaltensimplementierung

Grundlagen zum Korrektur- und Rücksprungverhalten

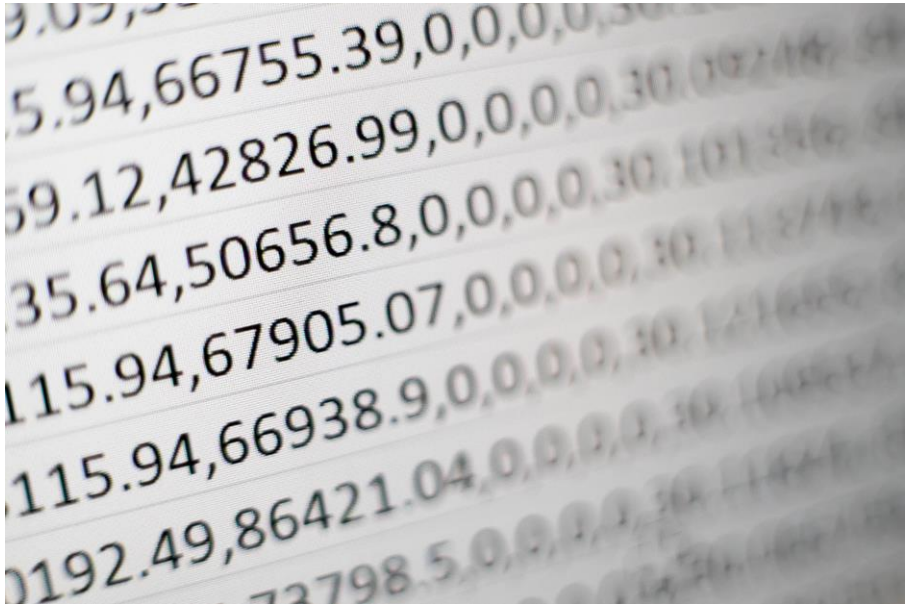
- explizit: Step-By-Step-Korrektur mit SI: „Korrektur"
- implizit: Korrektur-Oneshot mit „Nein ,ich meinte ..."
- Historischer vs. Hierarchischer Rücksprung

Disambiguierungen

- Taskdisambiguierung
- homophone/heterophone Disambiguierung
- phonetisch/textuelle Suchen

Allgemeine Auswahl aus Listen

- Listenarten: Statisch vs. Dynamisch
- Größe der Ergebnismenge: klein, groß und „unendlich"
- Zeilennummer anzeigen?
  - Keine Anzeige bei Assistenten im Smartphone → Erkannt wird sie trotzdem
  - Mit Anzeige und Erkennung im Auto
  - Zeilennummer muss immer spracherkennbar sein
  - 1. Zeile: die erste sichtbare Zeile im View oder die erste Zeile der Liste; Zeilenzahl wird am Ende der Seite weitergezählt
- statische Listen/Enumerationen
  - Zeilennummer & Zeileninhalt; bekanntes Set an Wörtern → Erkenner gut trainierbar
- große Ergebnismengen
  - Phoneme für sprechbare Zeileninhalt und angezeigte Zusatzinformationen müssen ggf. dynamisch erzeugt werden
  - PLZ, Orte, Alben, Künstler, Teile von Telefonnummern, Bewertungen u.a. Daten von Online-Diensten, …



Dialogmanagement / Verhaltensimplementierung

Applikationslogik

- Auflösen von Suchbereichen
- Umgang mit fehlende Daten
  - Bei Kontakten: Telefonnummern, Adressen, etc.
  - bei POIs: Online-Zusatzdaten
- Sonderfälle bei Car-Funktionen

Filtern von Ergebnissen

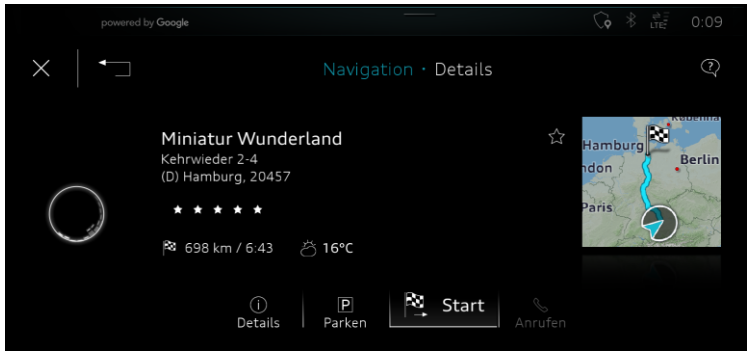
- welche Ergebnisse werden angezeigt?
- Kategorienübergreifen vs. Kategorienspezifisch

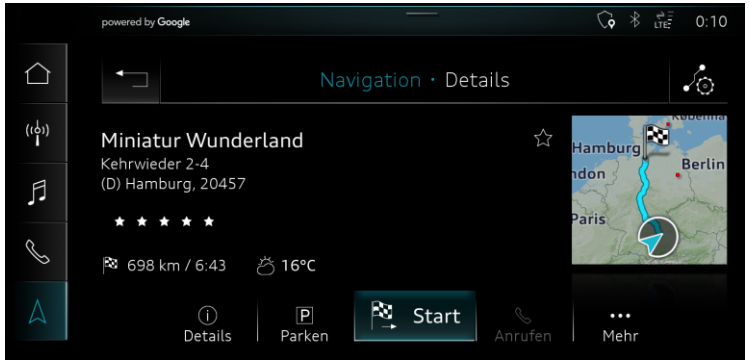
Sortieren von Ergebnissen

- in welcher Reihenfolge werden die Ergebnisse angezeigt?
- angewendetes Sortierkriterium anzeigen, sonst wird es nicht verständlich, wonach sortiert wurde
- Ergebnisse müssen (idealerweise sinnvoll) vergleichbar sein

Darstellungen auf Displays

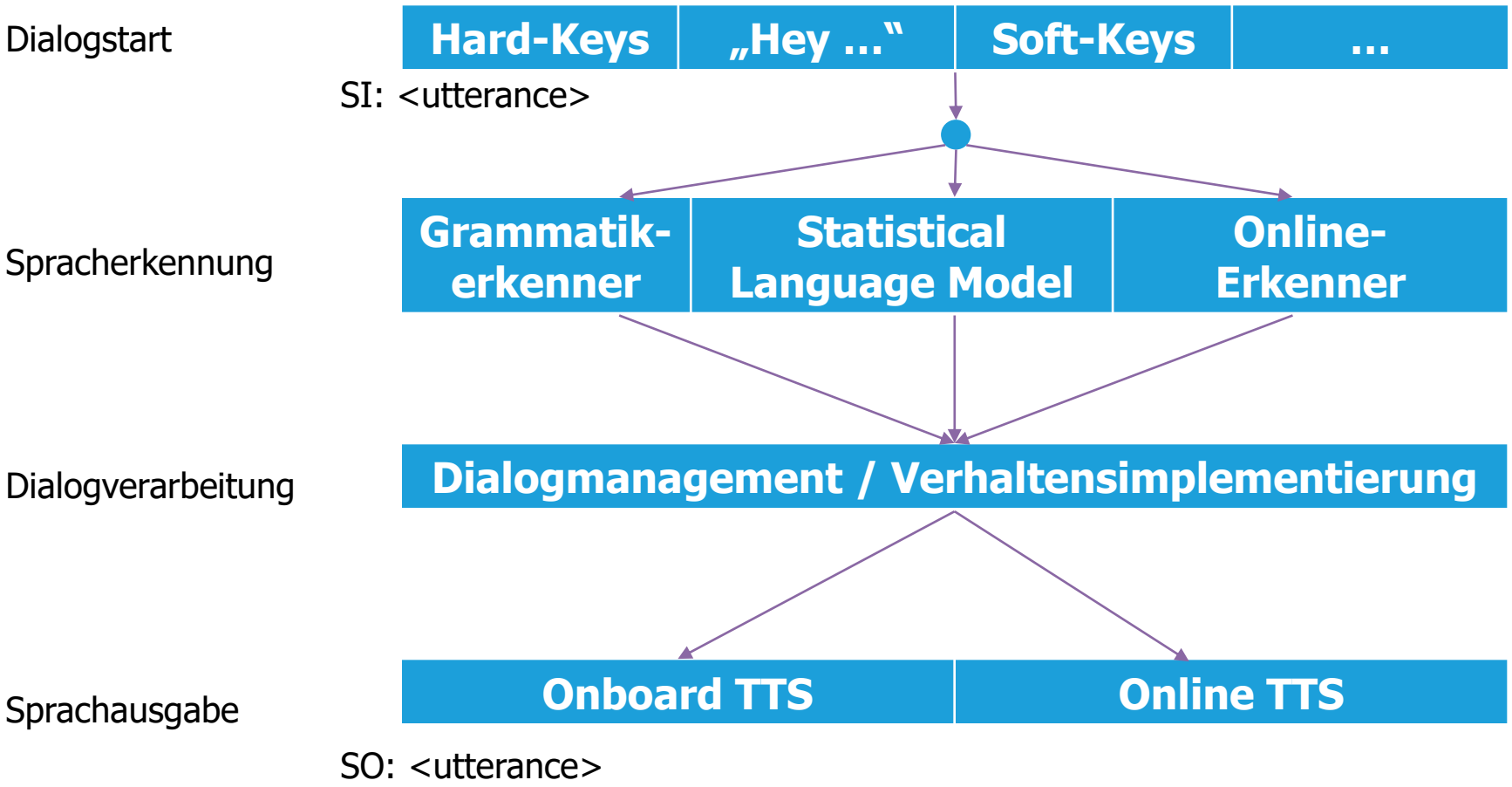
- Integration in bestehende GUI*
  - Nutzung bestehender GUI-Screens
  - SDS-relevante Elemente müssen integriert werden
- Eigene GUI der SDS
  - Läuft unabhängig von der GUI
  - Legt sich über die bestehende GUI
- In beiden Fällen:
  - Popups und Dialoge der GUI beachten





* GUI: „Graphical User Interface"; beschreibt die graphische Benutzungsschnittstelle und somit alles, was auf Displays ausgegeben wird.

Sprachausgabe



Sprachsynthese (TTS)

Onboard TTS:

- Verarbeitung/Aufbereitung passiert onboard
- Ausgabe passiert (auch) onboard
- Gut geeignet für Dinge, die immer und überall funktionieren müssen
- Behandlung von Sonderfällen (nur) nach Software-Updates
- funktioniert gut bei vorher definierten Texten (SDS-Prompts, Wetter-Dienst, Nav-Ansagen)
- Datenschutz beim Vorlesen von persönlichen Daten (SMS/E-Mails) ist gegeben

Online TTS:

- Verarbeitung/Aufbereitung passiert online
- Ausgabe passiert onboard
- Behandlung von Sonderfällen *jederzeit* machbar bzw. abh. Vom TTS-Anbieter
- Relativ robust gegen Sonderfälle wie „denglische" Wörter
- Gut geeignet für lange, dynamische (personenunabhängige) Texte mit unvorhersehbarem Inhalt

Ansprache und Stimmenauswahl

Ansprache: Duzen oder Siezen?

- Bisher: Siezen im Auto; Duzen in mobilen Assistenten
- Bonusfrage: Wer nimmt es dem System übler, wenn das falsche gewählt wird?
- Alternativ: unpersönliche Anrede

Männliche oder weibliche Stimme?

- kulturell abhängig
  - Arabien & Türkei: männlich
  - Der Rest weitgehend: weiblich (Alt)
- siehe Portfolio der Anbieter für Sprachsynthese

Promptformulierungen

Wie formuliert man Prompts?

- Vollständige Sätze, nicht nur kombinierbare Wortgruppen
- Möglichst in einer Sprache bleiben
- Problematisch sind fremdsprachige Begriffe („Business-Denglisch“)
- Deklination der Wörter beachten, vor allem bei dynamischen Anteile („… in <Country>“)

Bei Fremdsprachen

- Anzahl der Mehrzahlfälle sind sprachenabhängig
  - deutsch: 1 – viele
  - slawische Sprachen: 1 - 2..5 – viele
- Unterschiedliche Fälle bei bestimmten Wortarten abhängig vom Geschlecht
  - „dein Vater“ vs. „deine Mutter“
- Exonyme & Zeichensatz beachten
  - „Schloss Neuschwanstein“ – „Château de Neuschwanstein“ – „Замок Нойшванштайн“
  - „Собор Василия Блаженного“ – „Basilius-Kathedrale“

Pause-Verhalten

Automatische Pausenaktivierung:

- Bei eingehendem Telefonanruf
  - Bei angenommenen Telefonanruf: Dialogabbruch sinnvoll
  - Bei abgelehnten Telefonanruf: Dialogfortsetzung möglich
- Beim Scrollen in Listen
  - Auf Erkenner-Timeout achten
  - Ggf. Erkenner-Timeout verlängern/neusetzen oder den Pausemodus aktivieren
- auf Hilfeseiten mit viel Text
- aufgrund äußerer Umgebungsvariablen („Workloadmanager")

Explizite Pausenaktivierung:

- Druck auf Erkenner-Status - Play-Pause-Toggle
- SI: Pause

Grundlagen zum Korrektur- und Rücksprungverhalten

- Explizit: Step-By-Step-Korrektur
  - SI: „Korrektur"
  - 1 Schritt zurück
  - Letzte Frage/Prompt nochmal stellen
- Implizit: Korrektur-Oneshot
  - „Nein ich meinte ..."
  - 1 Schritt zurück direkt mit neuem Wert gekoppelt
- Historischer Rücksprung oder Hierarchischer Rücksprung
- Historischer Rücksprung kann beim Scrollen in Listen kontraproduktiv sein
- Hierarchischer Rücksprung berücksichtigt im Grundkonzept nicht die Disambiguierung

Dialogreaktivierung:

- Druck auf Erkenner-Status - Play-Pause-Toggle
- Druck auf PTT
- Druck auf Listenitem, mit direkter Zeilenauswahl/Screenbedienung und Dialogfortsetzung
- Nicht vergessen: Auch eine explizit aufgerufene Pause muss beendet werden!

Phonetische vs. textuelle Suchen

- Phonetische Suchen sind robust gegen:
  - Rechtschreibfehler
  - Unvollständige Suchen innerhalb eines dynamischen Anteils
  - Ungefähre Aussprache kann ausreichend sein
- Homophone Wörter sind auffindbar
  - M(a|e)(i|y)er, (Ch|K)ristin[e], ...
  - Lokale Adressbücher vs. Online-Suchen
- Phoneme notwendig
  - In Datenbanken vorhanden
  - Dynamisch erzeugt und als Guestkontext eingebunden
- Sonderfälle sind zu berücksichtigen
  - Abkürzungen/Sonderzeichen: z.B. P!nk, Toys'R'us, Inh., GmbH, DB (Deutsche Bahn)
  - Datenqualität
    - „Ikea in München-Eching“
    - „Firmenmuseum der Deutschen Bahn AG“ (aka „Eisenbahnmuseum“)

Disambiguierungsarten und deren Nutzung - Taskdisambiguierung

- Regelfall: Task wird über das Topic/Grammatikregel erkannt
- „Ich möchte <dynamischer Anteil>"
  - Kein Kontext erkennbar → je nach Art der dynamischen Anteile sind mehrere Kontexte denkbar
- „Ich möchte zu Müller"
  - Kontakt - anrufen, SMS/Email schreiben, hinfahren
  - Drogerie - hinfahren, anrufen
- Theoretisch jede Art eines dynamischen Anteils mit jedem Task im System kombinierbar → sehr komplex
- Praktischerweise: dialogseitig eingrenzen
- Funktioniert ggf. über Gruppen (Kontakte vs. POI, Kontakte vs. Künstler, Künstler vs. Album)

Disambiguierungsarten und deren Nutzung – Homophonie

- „Anrufen bei (Ch|K)ristin[e] M(a|e)(i|y)er“
  - Christin/Christine/Kristin/Kristine
  - Maier/Mayer/Meier/Meyer
- Zusatzdaten mit Anzeigen zur Unterscheidung
  - Bei exakt gleichen Kontakten/Daten Pflicht
- Auswahl über Zeilennummer ist Pflicht
- Idealerweise sollten die Zusatzdaten auch sprechbar sein
  - Kontakte
    - Email-Adressen → Achtung Sonderzeichen!
    - Telefonnummern, Parameter & auch Teile/Einzelziffern der Nummer
    - Firma → Abkürzungen berücksichtigen
  - Ziele
    - „Das in <PLZ, Bundesland, Stadt, Straße ...>“
    - Zusatzbezeichnungen („… an der Oder“, „… an der Weinstraße“)

Disambiguierungsarten und deren Nutzung – Heterophonie

Heterophone Disambiguierung

- Betrifft sehr ähnliche Orte (Neustadt, Neustedt)
- Häufigste Disambiguierung, weil auch allgemeine Auswahl aus Picklisten
- Verschiedene Items mit gleichen/ähnlichen Bezeichnungen
- Auswahl über Zeilennummer oder Zeileninhalt

Sonderfälle

- „Fahre mich nach Hause" vs. „fahre mich zu Herr Hausen"
- „Anrufen bei Mailbox" vs. „Mailbox anrufen"