Einführung in die Computerlinguistik

13: Verarbeitung gesprochener Sprache ("Speech Processing")

WS 2008/2009

Manfred Pinkal

Nachtrag zum Informations-Management: Informations-Extraktion und Question Answering

Informations-Extraktion

Who did what to whom?

- Fülle Rollen in Template mit Information
  - Ignoriere Rest des Textes
    - Information muß als Template darstellbar sein
    - Information muss mithilfe einfacher Regeln im Text identifizierbar sein
- Beispiele:
  - Vortragsankündigung (wer, wann, wo, worüber)
  - Wetterbericht (wann, wo, wie)
  - Wirtschaftsmeldungen (wer, wen, was)

Vortragsankündigung

Am Donnerstag, den 13.11.2006, redet Martha Palmer (University of Pennsylvania) um 16:15 im Seminarraum (Geb. 17.1) zum Thema „Putting Meaning into your Trees".

Redner: ?
Zeit: ?
Datum: ?
Ort: ?
Titel: ?

Schritt 1: Datenaufbereitung

- POS-Tagging
  - um, am, im: PRP
  - redet: VVFIN
- Morphologie

Einzelne Module entweder wissensbasiert oder statistisch

- Named Entity Recognition
  - PERSON, ORGANISATION, TIME, DATE, QUANTITY...
- Flache Grammatik
  - Phrasen erkennen
    - PRP + TIME → Präpositionalphrase (PP)

Schritt 2: Ereignismuster /“Event Patterns“

[Am DATE] redet PERSON (ORGANISATION) [um TIME] [im PLACE] [zum Thema [„Putting Meaning into your Trees“]].

- Regeln kodieren Wissen darueber, wie Information aus Template sprachlich ausgedrueckt wird („Abbildung Sprache nach Bedeutung“)
  - Wenn [$_{PP}$ um TIME], dann Zeit → TIME
  - Wenn [$_{PP}$ zum Thema S], dann Titel → S

Vortragsankündigung

Am [$_{PP}$ Donnerstag, den 13.11.2006], redet Martha Palmer (University of Pennsylvania) [$_{PP}$ um 16:15] [$_{PP}$ im Seminarraum (Geb. 17.1)] [$_{PP}$ zum Thema [„Putting Meaning into your Trees“]].

Redner: Martha Palmer
Zeit: 16:15
Datum: Donnerstag, den 13.01.2006
Ort: Seminarraum (Geb. 17.1)
Titel: „Putting Meaning into your Trees“

Beurteilung von Information Extraction

- Gut für Suche nach spezieller Information
  - Templates gut zur Weiterverarbeitung
  - Relativ sicheres Wissen
  - Einigermassen gut automatisierbar
- Problem: Flexibilität
  - Wortwahl: „über“ vs. „zum Thema“
  - Satzbau: „XY redet am 01.11“. vs. „Am 01.11. redet XY“
  - Abdeckung der Regeln?
  - Übertragbarkeit auf andere Domänen problematisch
- Rolle von sprachlichem Wissen:
  - Wissen ueber Domänenstruktur: Definition der Rollen
  - Sprachliche Realisierung von Rollen in Event Pattern

Question Answering

- Gegeben: Query (als umgangssprachliche Frage)
- Gesucht: Relevanter Satz (aus Dokument)

- Typische QA-Systeme machen nur Extraktion
  - Schritt 1: IR → Liste von Dokumenten
  - Schritt 2: Extraktion der relevanten Stellen

Zur Extraktion ist tiefe(re) Verarbeitung nötig!

Welche Stellen sind relevant?

- Zentrale Idee: Relevante Stellen treffen **Aussage** über das **gefragte Objekt**
  - Überlappung mit Frage in Worten ist zu unspezifisch
  - **Semantische** Repräsentation nötig
- Fragenklassifikation: Wonach wird gefragt?
  - „Wie viele Sechsecke sind auf einem Fußball?“ (Zahl)
    - `sein: ?<Zahl> Sechsecke, auf Fußball`
  - „Wo ging Bill Gates auf College?“ (Ort)
    - `gehen: Bill Gates, College, ?<Ort>`

- Strategie: Finde Aussage in Dokument, die die Lücke in der Anfrage füllen kann

Beurteilung von Question Answering

- Gibt relevanten Satz zurück
  - Benutzerfreundlichster Ansatz

- Question Answering ist schwer
  - Aufwändig
  - Robustheit großes Problem
  - Oft für begrenzte Domänen untersucht
    - Richtung „Expertensysteme“

- Rolle von sprachliches Wissen
  - Braucht **deutlich mehr** Wissen als reines Information Retrieval

Zusammenfassung und Ausblick

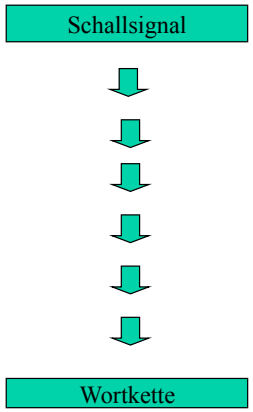
- Information Management ist schwierig
  - Wenig Wissen: erstaunlich gute Ergebnisse (IR)
  - Qualitativer Sprung (QA) erfordert viel Wissen

- Verschiedene Verfahren für verschiedene Aufgaben
  - Homogene Daten, kleine Domäne: Information Extraction
  - Domänenunabhängige Suche: Information Retrieval
  - Mit viel Wissen: Question Answering

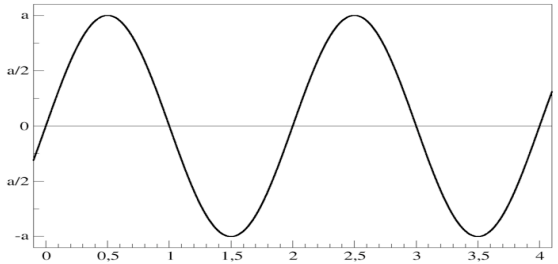
Spracherkennung

- Die Grundaufgabe der Spracherkennung: Gegeben ist ein kontinuierliches Schallsignal. Welche Kette von Wörtern wurde vom Sprecher geäußert?

Spracherkennung: (Vereinfachtes) Schema



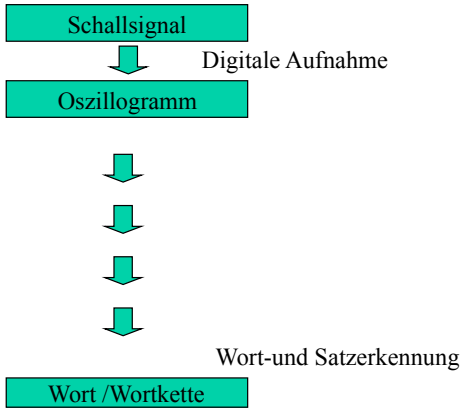
Reine Schwingung



Amplitude ≈ Schalldruck ≈ Lautstärke

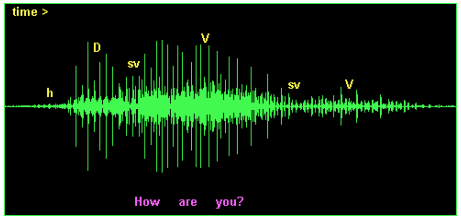
Frequenz ≈ Tonhöhe (engl. “pitch”)

Spracherkennung: (Vereinfachtes) Schema

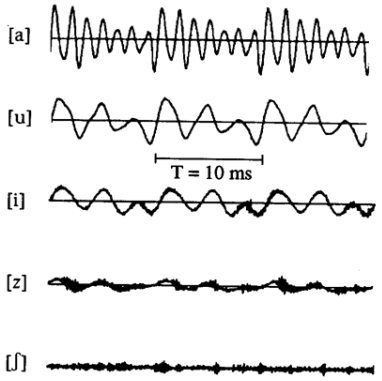


Realistisches Beispiel

- Das Oszillogramm für eine Äußerung des englischen Satzes „How are you“

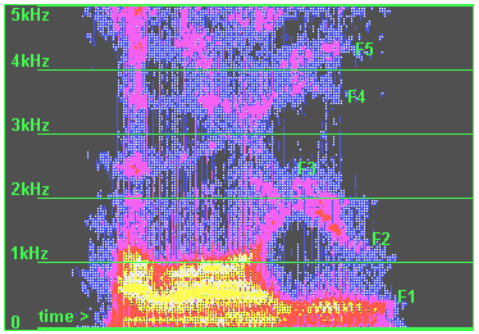


Einzelne Laute als Oszillogramme



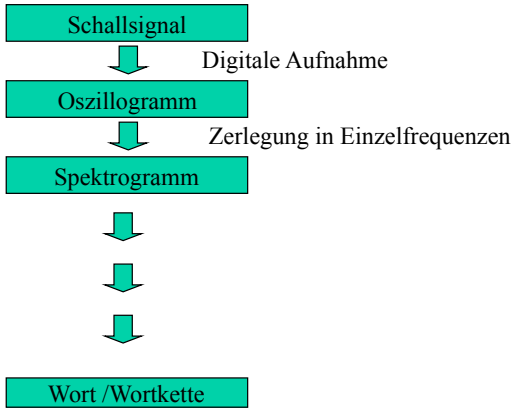
- Laute werden charakterisiert durch Kombination von Schwingungen verschiedener Frequenzen
- Im Oszillogramm **schwer erkennbar** (Überlagerung)
- Daher: Geschicktere Repräsentation durch Komponentenanalyse (Fourier-Transformation)
- Ergebnis: Zeit-Frequenz -Diagramm (**Spektrogramm**)

Ein Spektrogramm

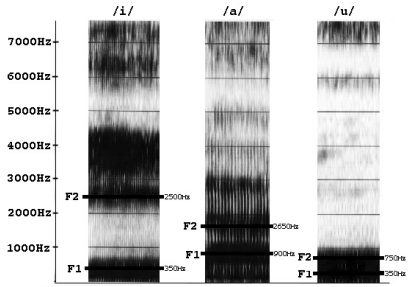


... für den englischen Satz „How are you?“

Spracherkennung: (Vereinfachtes) Schema

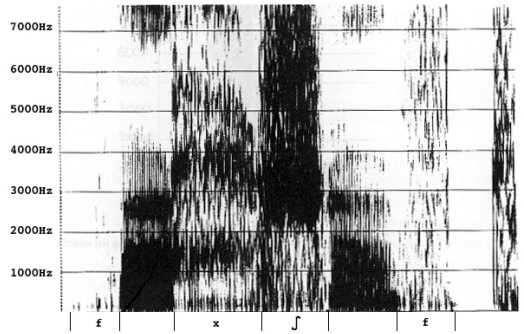


Spektrogramm für die Vokale i,a,u



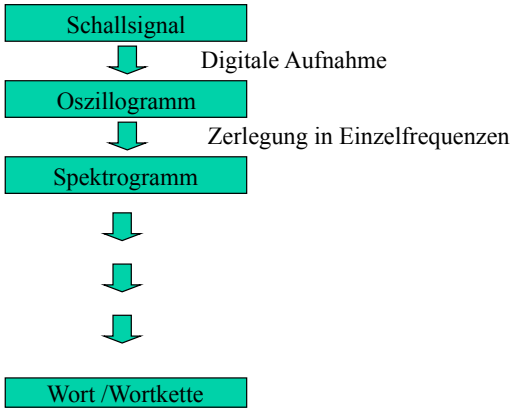
- Dunkle Färbung: große Schallenergie in einem bestimmten Frequenzbereich.
- Die Formanten (Obertöne) F1 und F2 sind für die charakteristische Vokalqualität verantwortlich.
- Der Verlauf des Basisformanten F0 (hier nicht sichtbar) gibt die Intonation der Äußerung wieder.

Spektrogramm für einige Konsonanten

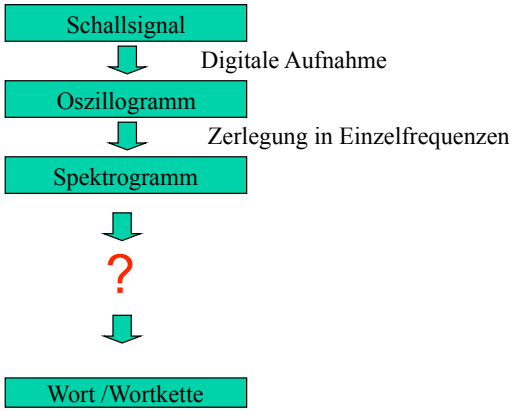


Frikative: f und ch-Laut („ach“-Laut); Sibillant: „sch“-Laut

Spracherkennung: (Vereinfachtes) Schema



Spracherkennung: (Vereinfachtes) Schema

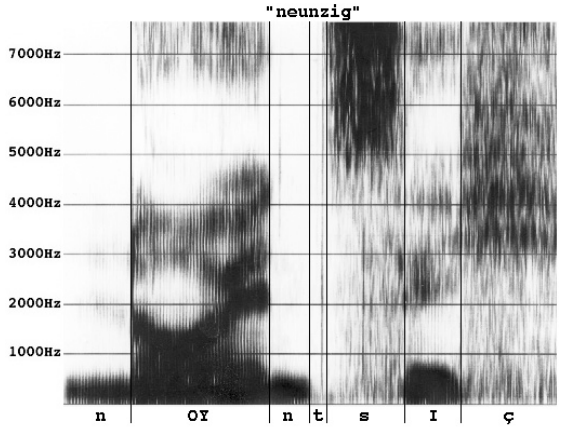


Naives Modell zur Lauterkennung

- Schritt 1: Identifikation einzelner Spektrogramm -Schnipsel = Laute (Segmentierung)
  - Finde “Übergänge” in Spektrogramm
- Schritt 2: Vergleiche Spektrogramm-Schnipsel mit Datenbank “idealer” Laute (Identifikation)
  - Identifiziere passende Phoneme
- Schritt 3: Setze orthographische Realisierungen der Phoneme hintereinander
  - Ergibt die entsprechenden Wörter

Funktioniert leider nicht!

Spektrogramm für ein deutsches Wort



Problem1: Kontinuität des Signals

- Die Laute eines Wortes lassen sich schwer abgrenzen
  - Wo hört Laut 1 auf, wo fängt Laut 2 an?
  - Dazu kommt das Phänomen der Koartikulation: Laute beeinflussen sich gegenseitig.
    - In Lautfolgen wie [am], [um], [an] kann man nicht den Vokal vom Nasal trennen: Vokal hat Nasal-Qualität und umgekehrt.
    - /k/ wird verschieden realisiert in Koffer, Kind, Kabel
- Wörter sind nur in der Orthografie sauber getrennt.
  - In der gesprochenen Sprache gibt es zwischen Wörtern meistens keine Pause
  - Pausen kommen in spontaner Sprache auch innerhalb von Wörtern vor

Problem2: Varianz der Realisierung

Gleicher Laut/ gleiches Wort wird nicht immer gleich ausgesprochen

- Verschiedene Dialekte
- Verschiedene Sprecher
- Unterschiedliche Sprechgeschwindigkeit
- Physischer und emotionaler Zustand des Sprechers
- Abhängig von Tonhöhe und Akzent

Problem3: Varianz des Signals

Sprachexterne Einflüsse verändern das Signal

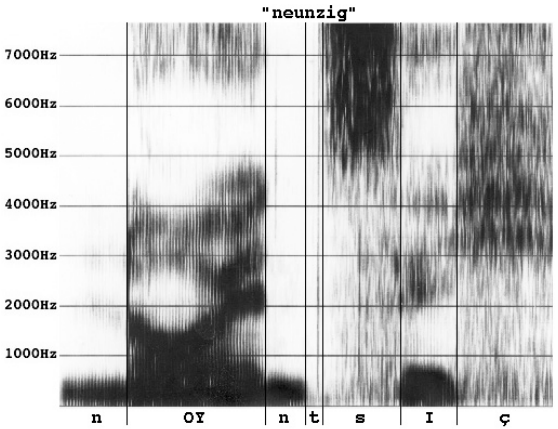
- Raumakustik, Hall, Entfernung
- Medium: Face-to-Face, Telefon, Handy
- Mikrofonqualität und -charakteristik
- Hintergrundgeräusche („Rauschen“, „Noise“)

Statistische Modellierung

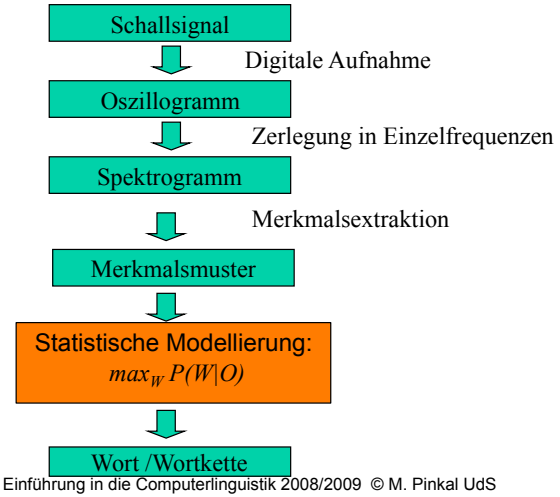
- Eine Art Klassifikationsaufgabe:
- Berechnung der wahrscheinlichsten Wortkette
  - $W = w_1 w_2 \ldots w_n$,

  die einem Eingabe-Ereignis O entspricht:
  - $max_W P(W|O)$ „das W, für das $P(W|O)$ maximal“
- Ermittlung der Eingabe-Ereignisse O aus dem Schallsignal durch Merkmalsextraktion:
  - Bestimmung der Schallenergie in einzelnen Frequenzfenstern (z.B. Viertelton) und Zeitfenstern (z.B. 20 ms).
  - Resultat ist eine Folge $O = o_1 o_2 \ldots o_m$
  - Jedes o_i ist ein Merkmalsvektor, der die Schallenergie für die unterschiedlichen Frequenzfenster in einem bestimmten Zeitfenster angibt.

Spektrogramm für ein deutsches Wort



Spracherkennung: (Vereinfachtes) Schema



Das Bayessche Theorem

- Wie bestimmen wir *P(W|O) = P($w_1 w_2$ ... w_n, o_1 o_2 ... o_m) ?*
- Das Bayessche Theorem oder die Bayes-Regel:

$$P(E \mid F) = \frac{P(F \mid E) \cdot P(E)}{P(F)}$$

- Die Bayes-Regel ist ein elementares Gesetz der Wahrscheinlichkeitstheorie. Sie ist überall da nützlich, wo der Schluss von einer Größe F auf eine andere Größe E bestimmt werden soll (typischerweise von einer Wirkung/ einem Symptom auf die mögliche Ursache), die Abhängigkeit in der anderen Richtung (von der Ursache auf die Wirkung) aber besser zugänglich ist.

Das Bayessche Theorem

- Wie bestimmen wir *P(W|O) = P($w_1 w_2$ ... w_n, o_1 o_2 ... o_m) ?*
- Grundaufgabe der Spracherkennung: Wie schließen wir vom akustischen Merkmalsmuster *O* (dem Symptom) auf die Wahrscheinlichkeit der tatsächlich geäußerten Wortkette *W* (die Ursache)?
- Bayes-Regel :

$$P(W \mid O) = \frac{P(O \mid W) \cdot P(W)}{P(O)}$$

- Die wahrscheinlichste Wortkette:

$$\max_W P(W \mid O) = \max_W \frac{P(O \mid W) \cdot P(W)}{P(O)}$$

$$= \max_W P(O \mid W) \cdot P(W)$$

Akustisches Modell und Sprachmodell

$$\max_W P(W \mid O) = \max_W P(O \mid W) \cdot P(W)$$

- *P(O|W)* ist die Wahrscheinlichkeit, dass eine Wortfolge in einer bestimmten (durch den Merkmalsvektor bezeichneten) Weise ausgesprochen wird: Akustisches Modell
- *P(W)* ist die Wahrscheinlichkeit, dass eine bestimmte Wortfolge geäußert wird: „Sprachmodell“

Akustische Modelle

- Training von „Lautmodellen“: Aufnahmen von Sprachlauten mit ihrer phonetischen Kategorie/ Umschrift
- Aussprachewörterbuch, das für jedes Wort die phonetische Umschrift enthält
  - Genauer: Die Umschrift für alternative Aussprachen, die in einem gewichteten endlichen Automaten kodiert sind.
- Für die statistische Zuordnung von Merkmalsmustern und Wörtern wird die HMM-Technik („Hidden Markov Models“) verwendet.

Sprachmodelle

- Wie berechnen wir $P(W) = P(w_1 w_2 \ldots w_n)$?
- Grundlage ist die Frequenz von Wortfolgen in Korpora.
- Sparse-Data-Problem: Ganze Sätze kommen viel zu selten vor.
- Kettenregel erlaubt die Reduktion von $P(w_1 w_2 \ldots w_n)$ *auf bedingte Wahrscheinlichkeiten:*

  $P(w_1 w_2 \ldots w_n)$

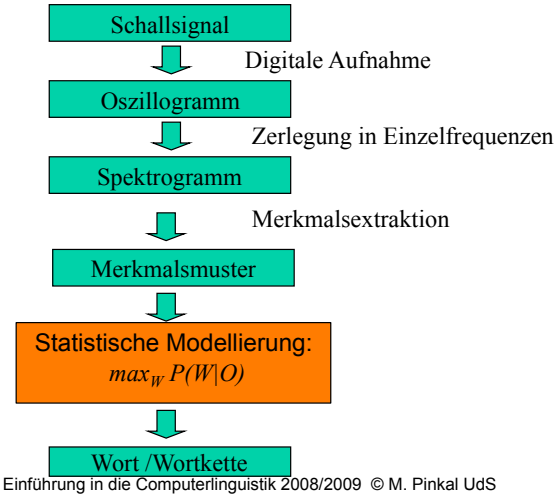
  $= P(w_1) * P(w_2|w_1) * P(w_3|w_1 w_2) * \ldots * P(w_n|w_1 w_2 \ldots w_{n-1})$
- $P(w_n|w_1 w_2 \ldots w_{n-1})$*:* Sparse-Data-Problem ist nicht beseitigt!

Sprachmodelle

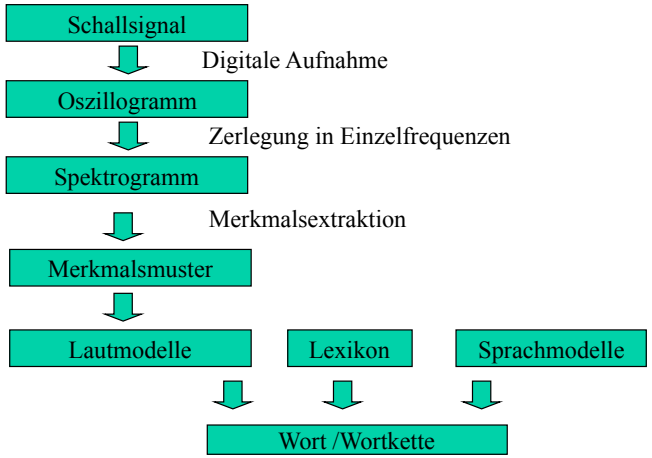
- n-Gramm-Technik: Wir approximieren die Wahrscheinlichkeit, dass ein Wort w im Kontext einer beliebig langen Wortfolge auftritt, durch die relative Häufigkeit, mit der es in einem auf n Wörter begrenzten Kontext auftritt.
  - Dabei wird das Wort selbst mitgezählt. N-Gramm -Wahrscheinlichkeit berücksichtigt also einen Vorkontext von n-1 Wörtern.
- Meistens wird mit Bigrammen und Trigrammen gearbeitet.
- Bigramm-Approximation:
  - $P(w_n|w_1 w_2 \ldots w_{n-1}) \approx P(w_n|w_{n-1})$

    $P(w_1 w_2 \ldots w_n) \approx P(w_2|w_1) * P(w_3|w_2) * \ldots P(w_n|w_{n-1})$

Spracherkennung: (Vereinfachtes) Schema



Spracherkennung: Schema

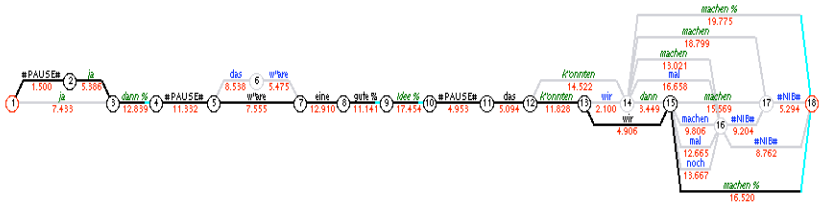


Erkennerausgaben

- Die „beste Kette“ (oder die n besten Ketten), ggf. mit „Konfidenzwert“ (einem Maß für die Verlässlichkeit der Hypothese).

- Alternativ: Ein Worthypothesengraph: Auf der Zeitachse werden die „geratenen“ Wörter mit ihrem zugehörigen Zeitintervall und einem Wahrscheinlich -keitswert abgetragen.

Ein Worthypothesengraph (WHG)



Quelle: Verbmobil, Terminvereinbarungsdialoge:

„*Ja, das wäre eine gute Idee. Das könnten wir dann machen*“