

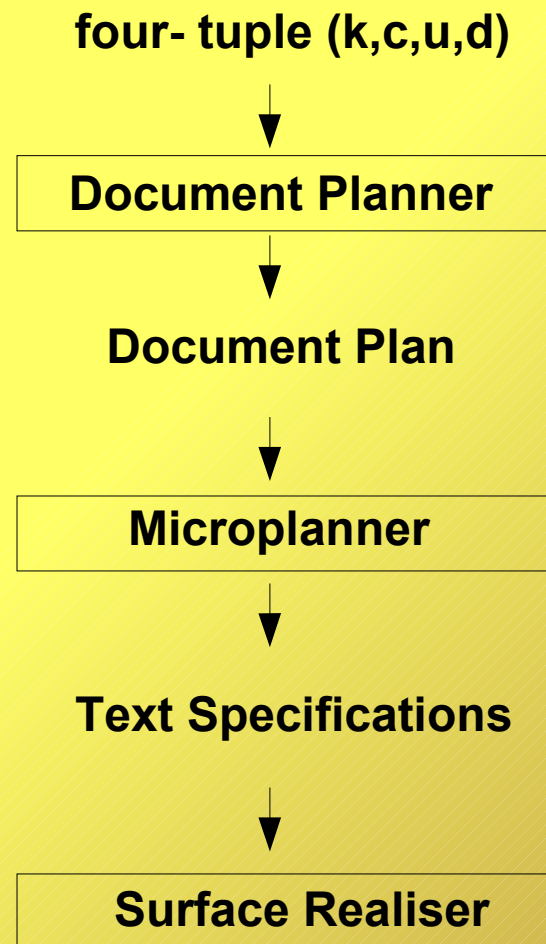
Architektur eines NLG- Systems

Tim Höcker

3 Phasen

1. Dokumentplanung (document planner)
2. Mikroplanung (microplanner)
3. Oberflächenrealisation (surface realiser)

Aufbau des Generierungssystems



Four-Tuple (k,c,u,d)

- k = Knowledge Source
- c = Communicative Goal
- u = User Model
- d = Discourse History

Wissensquelle (Knowledge Source)

- enthält Wissen über die Domäne, encodiert in Daten- und Wissensbasen
- anwendungsabhängig
- daher keine charakteristische Standardscheinung

Communicative Goal

- nicht zu verwechseln mit der *overall purpose*
- definiert die kommunikative Zielsetzung eines Textes
- verschiedene kommunikative Ziele (x,y,z,)
- wenn x mit einem Wert aus der Datenbank gefüllt wird, dann spricht man von einer Instanz eines kommunikativen Ziels

User Model

- charakterisiert den Benutzer
- Domänenwissen
- Technisches Vokabular
- stilistische und linguistische Vorlieben
z.B. *it's* vs. *it is*
- legt aufgrund dieser Charakterisierung fest, welche Informationen betont oder weggelassen werden sollten

Discourse History

- verantwortlich für die korrekte Benutzung anaphorischer Stilmittel z.B. Pronomen
- kann Vergleiche zu vorherigen Aktionen zeigen
- *kann* null sein

Document Planner

- Content Determination
- Document Structuring

Content Determination

- welche Informationen sprachlich im fertigen Text generiert werden sollten
- Faktoren, welche die Inhaltsfestlegung beeinflussen:
 - Charakteristika des Benutzers (Dau)
 - unterschiedliche *communicative goals*
 - *Constraints !?!*
 - erwähnenswerte Besonderheiten in der Informationsquelle
- anwendungsabhängig, daher keine Festlegung von generellen Regeln möglich
- GRICEsche Maxime

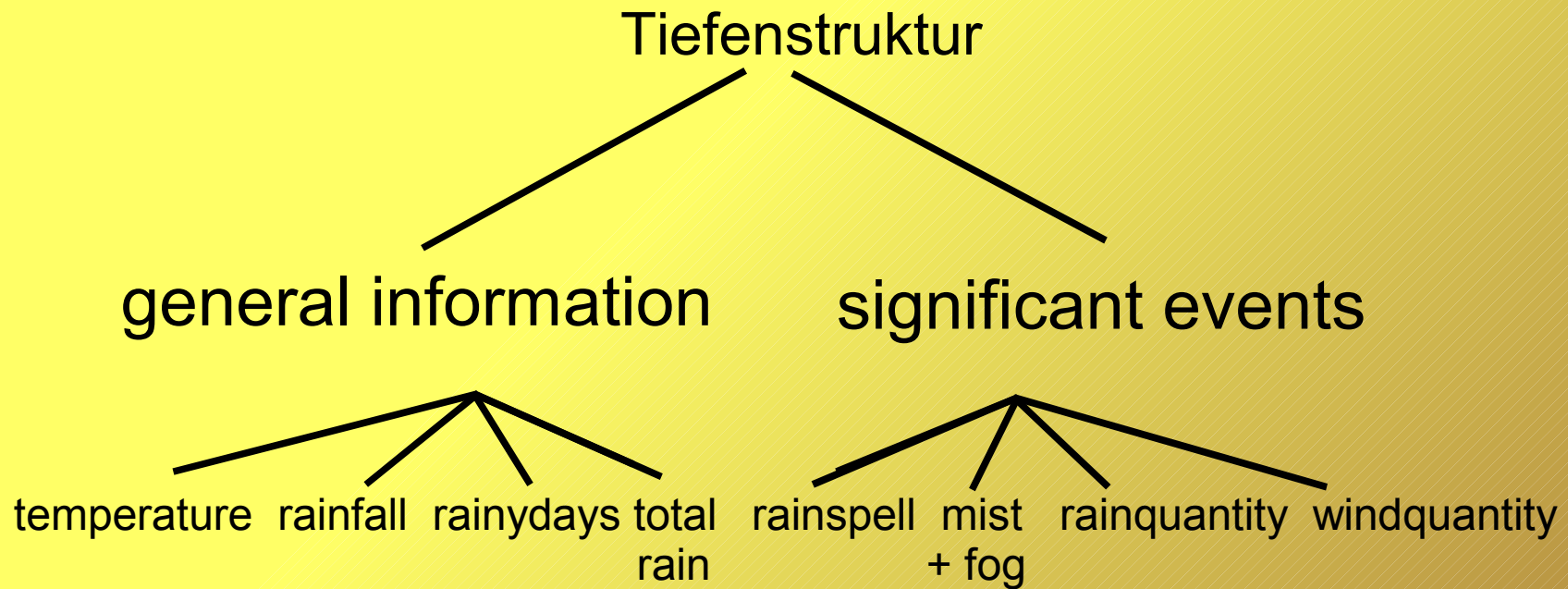
Document Structuring

- den zu übermittelnden Daten wird eine Struktur auferlegt
- Rezipienten haben genreabhängige Erwartungen an einen Text
- beinhaltet mehr als nur das Ordnen der Informationen an der Oberfläche
- Tiefenstruktur

Tiefenstruktur

1. alle Informationen werden ihrem Inhalt entsprechend geordnet, z.B. Temperaturinfos in einen Satz, alle Niederschlagsinfos in den nächsten Satz, usw.
2. Informationen werden unter Berücksichtigung der Verlaufsrelation gruppiert, z.B. allgemeine Aussagen über das Wetter in einen Abschnitt, detailliertere Aussagen in den nächsten Abschnitt

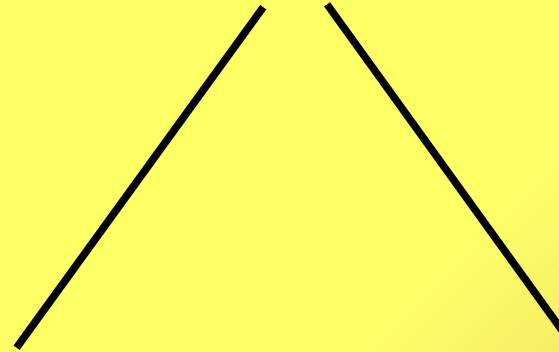
The month was cooler and drier than average, with the average number of rain days. The total rain for the year so far is well below average. There was rain on every day for eight days from the 11th to the 18th, with mist and fog patches on the 16th and 17th. Rainfall amounts were mostly small, with light winds.



Document Plan

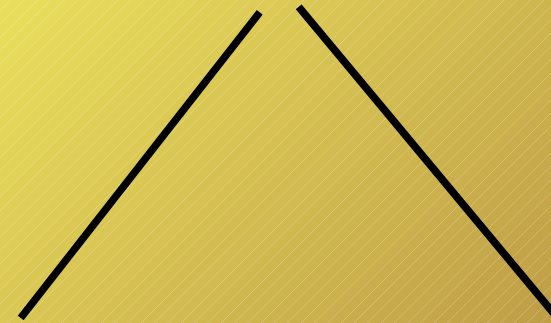
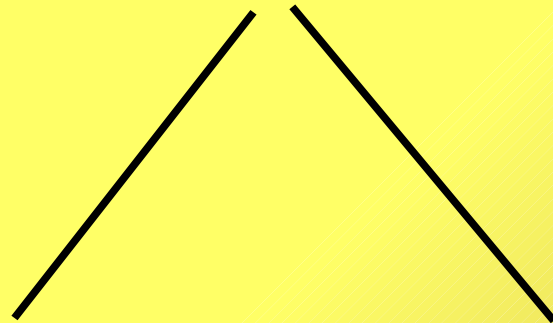
- Output of the Document Planner
- keine komplette Spezifizierung
- viele verschiedene Texte wären kompatibel mit den Informationen des Document Planner, die Feinabstimmung erfolgt in der Phase des *Microplanning*

Document Plan



Strukturinformationen

Discourse Relations



Inhaltsangaben Messages

Inhaltsangaben Messages

Messages

- im einfachsten Fall korrespondiert eine Message mit einem Satz
- abstrakte Informationen, die zwischen den Datenstrukturen des Anwendungsprogramms und dem zu generierenden Text vermitteln
- Hierfür muss eine domänspezifische Konzeptrepräsentation erstellt werden, d.h. Es müssen Entitäten, Konzepte und Relationen für bestimmte Domänen definiert werden. Wie das geht erfahren wir in der nächsten Stunde

Microplanning

- Lexicalisation
- Referring Expression Generation
- Aggregation

Lexicalisation

- wählt die adäquaten Wörter, um den zuvor ausgewählten Inhalt wiederzugeben
- entscheidet zwischen verschiedenen linguistischen Realisierungsmöglichkeiten
z.B. *the car owned by Mary* vs. *Mary's car*
- Problem: Multilinguale Programme. Man könnte unabhängige Lexikalisierungssysteme für die Zielsprachen konstruieren, aber *wasteful*. Ideale sprachunabhängige Regeln z.B. 'vermeide unnötige lange Auflistung', sollten über verschiedene Sprachen verteilt gelten, aber wie das funktionieren soll steht in den Sternen.

Referring Expression Generation

- verantwortlich für die eindeutige Identifizierung von Entitäten
- Erster Bezug: Initial Reference
- Zweiter Bezug: Subsequent Reference (oft Pronomen)
- Ambiguität muss vermieden werden
- Beispiele:
 - 1. März 2003
 - März
 - März vorigen Jahres
 - letzten Monat
 - Es

Aggregation

- entscheidet über die linguistische Struktur, mit der die zuvor bestimmten Infos generiert werden (Absätze, Sätze, Nominalphrasen)
- Sätze werden hinsichtlich ihrer Reihenfolge geordnet
- Interaktion mit *Lexicalisation*
- Eine Verbesserung des Flusses und der Lesbarkeit eines Textes sollte ausschlaggebend für die verwendete Ausdrucksweise sein.

Beispiel Satzaggregation

- Der Monat war kühler als der Durchschnitt.
- Der Monat war trockener als der Durchschnitt.
- Der Monat war kühler und trockener als der Durchschnitt.
- Der Monat war kühler als der Durchschnitt, mit weniger Niederschlag als gewöhnlich.

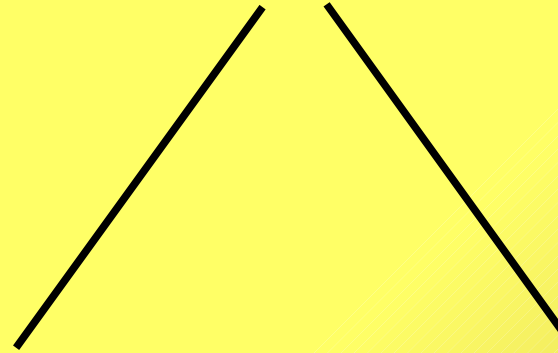
Microplanning

- Document Planner: Die Info, dass ein Zug um 7:40 ankommt und um 7:45 Richtung Köln abfährt, ist relevant und soll generiert werden.
- Microplanner: Wie soll das ausgedrückt werden.
- Es wäre möglich die Phase des Microplanning zu überspringen und alle Entscheidungen bereits in der Phase des Documentplannings zu treffen. Die Messages enthielten dann konkrete Informationen, die jeweils in einem einzelnen Satz ausgedrückt würden und somit könnte z.B.auf Aggregation verzichtet werden. Dies würde allerdings eine Einschränkung der Flexibilität bedeuten.

Text Specifications

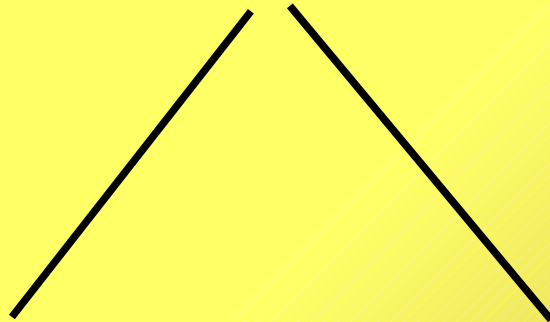
- Output of the Microplanner
- enthält alle Daten, die für die Generierung notwendig sind
- interne Knoten spezifizieren die Struktur (Absätze, Überschriften, usw...)
- Blätter sind *Phrase Specifications*
- Mehrere Messages können zu einer Phrase Specification zusammengefasst werden

Text Specifications



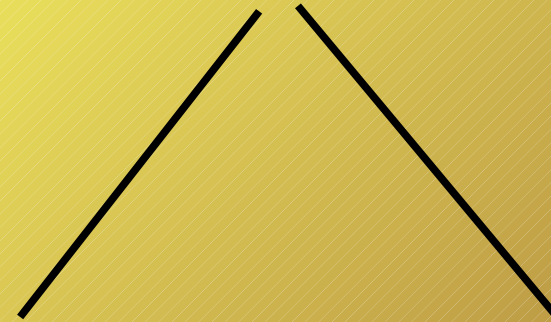
Strukturinformationen

Strukturinformationen



Phrase
Specifications

Ph Spec



Ph Spec

Ph Spec

Phrase Specifications

- spezifiziert einen Satz oder eine Satzeinheit
z.B. NP bei Überschrift
- kann in verschiedenen Graden der Abstraktion auftreten
- Orthographic String
- Canned Text
- Abstract Syntactic Structure
- Lexicalised Case Frames
- je abstrakter, desto komplexer, desto flexibler

Orthographic String

```
type: PSOrthographicString  
body: The month had some rainy days.
```

Canned Text

type: PSCannedText

text :|the month had some rainy days|

Abstract Syntactic Structure

type: PSAbstractSyntax

head: |have|

features: [tense: past]

subject: [type: PSAbstractSyntax
head: |month|

features: [definite: true]

[type: PSAbstractSyntax
head: |day|

object: [features: [number: plural

definite: false]

[type: PSAbstract Syntax

modifiers:

[head: |rainy|]

Lexicalised Case Frames

- Konstituenten werden durch semantische Rollen spezifiziert
z.B. *possessor* oder *possessed*
- sem. Kategorisierung hat gegenüber der syn. Kategorisierung den Vorteil, dass der Benutzer mehr Spielraum hat

Surface Realisation

- strukturelle Realisation
- linguistische Realisation

Strukturelle Realisation

- die meisten NLG-Systeme benutzen ein externes Präsentationssystem (Netscape, Word)
- für die internen Knoten (Struktur), müssen entsprechende mark-up-Symbole generiert werden
z.B. <p> oder \par

Linguistische Realisation

- Input sind die Text Specifications
- müssen in einen Text konvertiert werden
- verantwortlich für die grammatische Wohlgeformtheit
z.B. korrekte Wahl der Verbflexion
- etablierteste Bereich der NLG
 - REALPRO: Abstract Syntactic Representations
 - SURGE, KPML: Lexicalised Case Frames