Lexikonaufbau und Morphologieanalyseverfahren: Lexikalische Semantik

Gerold Schneider

Institut für Computerlinguistik, Uni Zürich

- Inhalt Teil 1: Wortbedeutung
  - ⟶ Bedeutung, Lexikon, Sinnrelation, Wortnetze
  - ⟶ Semantische Rollen, Selektionsrestriktion
  - ⟶ Das Generative Lexikon
- Inhalt Teil 2: Wortvorkommen
  - ⟶ Type/Token, Zipf
  - ⟶ Konkordanzen, Kollokationen
  - ⟶ Lexikologie – Lexikographie
  - ⟶ Korpuslinguistik, Empirie

TEIL I: Wortbedeutung

- Denotat: das von einem Wort (Wortgruppe) bezeichnete Ding (im weitesten Sinne)
- 'der Aussenminister von Polen' bezeichnet eine bestimmte Person
- Sinn und Bedeutung (klassisch: Frege)
  - Bedeutung: das Denotat, Sinn: die Art des Gegebenseins
  - Abendstern - Morgenstern: gleiche Bedeutung (im Fregeschen Sinn), die Venus
  - Art des Gegebenseins: Planet der am Abend/Morgen ....
- Freges 'Bedeutung' ist extensionale Bedeutung
- Freges 'Sinn' ist intensionale Bedeutung
- Extensional: 'Mensch' = Menge aller Menschen, 'Mars-Mensch' hat keine Bedeutung
- Intensional: 'Mensch' = Menge der definierenden Eigenschaften, die es erlauben, ein Exemplar der durch 'Mensch' denotierten Menge zu identifizieren

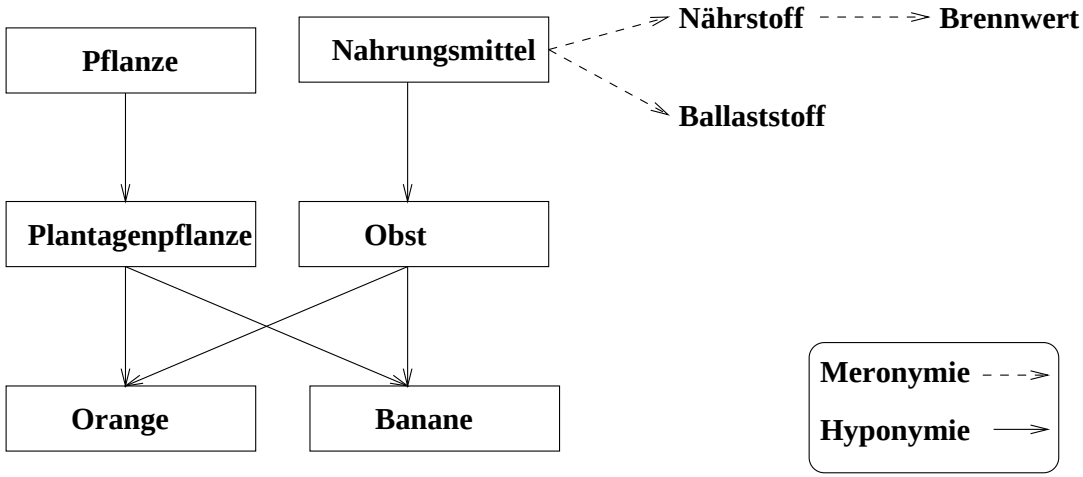
Bedeutung, fortg.

- Achtung: wir verwenden den Begriff Bedeutung *nicht* im Fregeschen Sinne!
- Alternative Konzeption: Bedeutung als mentales Objekt
- Alternative Konzeption: Bedeutung als repräsentationales Objekt (das Konzept, der Begriff)
- Worte und ihre Bedeutungen sind nicht isoliert, sie stehen in Beziehung (Sinnrelationen)
  - weiss - schwarz, Forelle - Fisch, Baum - Ast, ...
- Wortbedeutung ist oft schwammig
  - Vagheit: Onkel (Bruder der Mutter oder des Vaters)
  - Bedeutungsvielfalt: Schloss (Türschloss, Königsdomizil)
- Bedeutung vielleicht (oft) sowieso nur kontextuell gegeben ...
- Ist die Bedeutung von Wörtern atomar? Bedeutungsdekomposition
  - rennen = schnelles Fortbewegen, schlendern = gemächliches Fortbewegen

Struktur des Lexikons

```
beschaffen:
CAUSE(ACT(r),
      ET(BEC(HAVE(q,u)),
         BEC(NOT(HAVE(p,u)))
      ))

vorenthalten:
CAUSE(BEH(p),
      NOT(BEC(HAVE(q,u))))
```



- das Lexikon ist strukturiert:
  - Wörter haben interne Struktur (kompositionale Betrachtungsweise) → generative Semantik, Jackendoff'sche Semantik, CD-Theorie von Schank, Emphasetheorie nach Kunze, der frühe Rumelhard
  - Beziehungen zwischen Wörtern (relationale Betrachtungsweise) → Sinnrelationen oder lexikalische Funktionen, z.B. in WordNet
  - WordNet: http://www.cogsci.princeton.edu/ wn/wn1.7.1.shtml

Bedeutungsdekomposition: Emphasetheorie

- Definition einer finiten Menge primitiver Prädikate, aus denen sich die Wortbedeutung kompositional zusammensetzen lässt (töten als CAUSE TO DIE)
- Bsp. Emphasetheorie:

```
beschaffen:
CAUSE(  ACT(r),
        ET(  BEC(HAVE(q,u)),
             BEC(NOT(HAVE(p,u)))))

vorenthalten:
CAUSE(  BEH(p),
        NOT(BEC(HAVE(q,u))))
```

- z.B. vorenthalten: p verhält sich so (BEHAVE), das q nicht bekommt (BECOME) haben (HAVE) u
- Quelle: http://www.darmstadt.gmd.de/KONTEXT/TELEX/

Literatur: Kunze, Jürgen (1994): Verbfeldstrukturen. Berlin: Humboldt-Universität (http://www.darmstadt.gmd.de/KONTEXT/TELEX/Literatur/Kunze-Verbfeldstrukturen.pdf)

Bedeutungsvielfalt: lexikalisch - konventionell

- Polysemie: ein Lexem hat verschiedene Bedeutung (die sich auf eine Grundbedeutung zurückführen lassen)
  - gehen: fortbewegen versus funktionieren
- Homonymie: verschiedene Lexeme mit je eigenen Bedeutungen
  - modern: zeitgemäss versus verschimmeln
- Unterscheidung oft nur etymologisch begründbar und schwer zu unter- /entscheiden
- Spezialfälle der Homonymie:
  - Homographie:modern (i.G. zu altmodisch) vs. modern (i.S.v. verschimmeln)
  - Homophonie: Meer vs. mehr

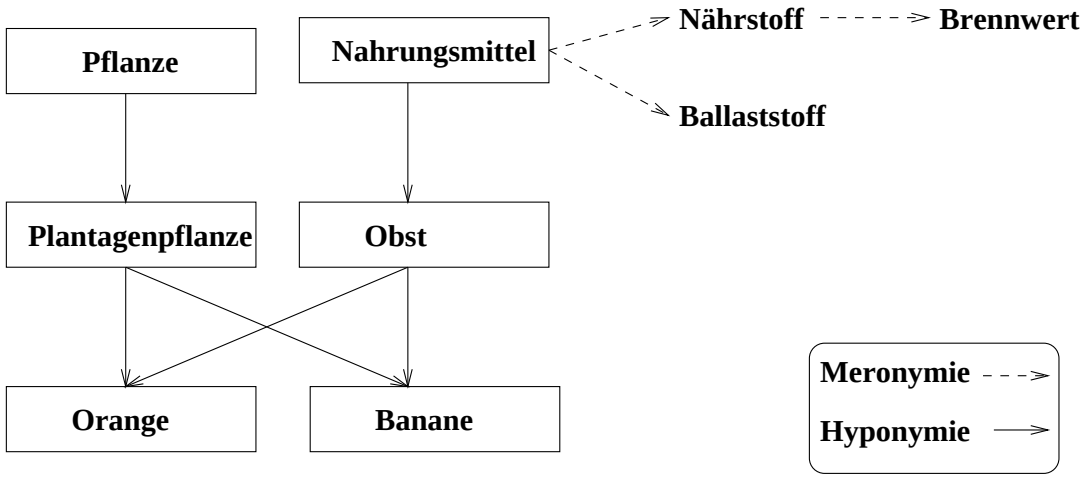
Bedeutungsvielfalt: figurativ - situativ

- Metapher: Ähnlichkeitsbeziehung: 'Theo ist (wie) ein Löwe'
- Metapher: 'Die Sonne lacht' im konventionalisiert-metaphorischen Sinne,
  → 'lachen' bedeutet nicht z.B. 'Sonnenstrahlen aussenden'
- Metonymie: Kontiguitätsbeziehung: 'Der neue (Computer der Firma) Apple ist grandios'
- Metonymie: 'Bern hat die Steuern erhöht'
  → 'Bern' sollte im Lexikon nicht die Bedeutung von 'Schweizer Regierung' bekommen
- Metonymie: 'Das Schnitzel an Tisch drei will zahlen'
  → 'Schnitzel' sollte im Lexikon nicht die Bedeutung von 'Gast' bekommen
- konventionelle Bedeutungsaspekte von Wörtern: vorhersagbare Polysemie
  → Schule als Gebäude, Summe von Ereignissen, Bildungsinstitut, ... (vgl. Bierwisch)

Sinnrelationen, lexikalische Funktionen

- Sinnrelation: Beziehung zwischen Wortbedeutungen
- Konnotation: nicht-definitorischer, assoziativer Bestandteil der Bedeutung (z.B. gefühlsbezogen)
- Synonymie: gleiche Bedeutung, unterschiedliche Konnotation. Synonyme sollten in jedem erdenkbaren Kontext austauschbar sein (Wahrheitswert wird erhalten)
  - Herstellung - Produktion - Fertigung
  - sterben - den Löffel abgeben - abkratzen - dahinscheiden - über den Jordan gehen
- Plesionymie: haben ähnliche, aber nicht identische Bedeutung
  - Geld - Kies
  - Auch feinere Charakterisierungen möglich: Emphase (Gegner vs. Feind), Implikation (verlegen vs. verlieren), Register (betrunken vs. sternhagelvoll)

Sinnrelationen fortg.



- Hyponymie/Hyperonymie: Subklassenbeziehung zwischen Lexemen (Konzepten?).
  - Raubvogel ist Hyponym von Vogel
  - Vogel ist Hyperonym von Raubvogel
- Auf der Konzeptebene spricht man hier von Ober-Unterbegriffsrelation, oder ISA-Hierarchie
- die so entstandene Hierarchie wird Taxonomie genannt (als Bestandteil einer Ontologie)

Sinnrelationen fortg.

- Holonymie/Meronymie: Teil-Ganzes-Relation, 7 Formen
  - Objekt-Komponente (Tasse vs. Henkel)
  - Ereignis-Merkmal (Rodeo vs. Cowboy)
  - Menge-Mitglied (Wald vs. Baum)
  - Masse-Portion (Brot vs. Brotscheibe)
  - Prozess-Phase (Heranwachsen vs. Kindheit)
  - Gebiet-Ort (Raum vs. Fenster)
  - Objekt-Material (Bierglas vs. Glas)
- Inkompatibilität
  - Antonymie: Gegensatz (hell-dunkel) - graduierbares Feld als Kriterium
  - Komplementarität: Ausschluss (Norden - Süden)
  - Konversion: verteile Rollenverhältnisse, die symmetrisch ausgetauscht werden können (kaufen - verkaufen)

Konzept vs. Wort/Lexem

- Konzept als sprachunabhängiger Repräsentant von Wörtern
- Konzept als Repräsentant von Synonymen
- Sinnrelation eigentlich (Abstraktionen) konzeptuelle Relationen?
  - Das Wort ’Haus’ hat-teil ’Fenster’ oder das Konzept ’Haus’ hat-teil ’Fenster’?
- Type - Token (abstrakte Wortform ist Type, Mehrfachauftritte im Text sind Token)
- Instanz (Exemplar, Individuum) - Konzept (Klasse von Individuen, Instanzen)
- Denotat/(Diskurs-)Referenz/Instanz: das (in der Welt, im Diskurs) mit einem Wort Bezeichnete

WordNet

- WordNet ist psycholinguistisch motiviert: Modell der Struktur des menschl. lex. Gedächtnis
- rekonstruiert Beziehungen zwischen Wörtern anhand von Sinnrelationen
- Organisation anhand taxonomischer Relation (Hyponymie/Hyperonymie), Verbindungen (Wortassoziationen?) mittels Sinnrelationen (semantische Nähe damit definierbar)
- fünf Kategorien: Nomen, Verben, Adjektive, Adverbien und Funktionswörter (noch nicht)
- Wörter werden auch glossiert, Nähe zu Thesaurus
- WordNet, EuroNet, GermaNet, UniNet ....
- Details siehe nächste Woche

Semantische Netze: etwas formaler

- Menge der Objekte O, alle Beziehungen sind 2-stellige Relationen
- Hyperonymie als 2-stellige Relation $ISA \subset O \times O$. ISA ist transitiv, asymmetrisch
  - die transitive Hülle ISA^* der ISA-Relation ordnet jedem Objekt alle seine Hyperonyme zu (die natürlich auch mittels einer rekursiven Suchprozedur berechnet werden können)
- Meronymie als 2-stellige Relation $MERO \subset O \times O$ (transitiv, asymmetrisch)
  - die transitive Hülle $MERO^*$ der $MERO$-Relation ordnet jedem Objekt alle seine Einzelteile zu (also auch die Einzelteile der Einzelteile)
- usw. für andere Relationen
- Was ist Vererbung formal? Die Komposition der transitiven Hülle der ISA-Relation mit der $MERO$-Relation ...
- vererben sich alle Sinnrelationen?

Konzeptdefinition in terminologischen Logiken (TL)

- Terminologische Logiken (auch Beschreibungslogiken) als Weiterentwicklung von semantischen Netzen (R. Quillian, 1969) und Frames (M. Minsky, 1974)
- KL-ONE (R.J. Brachman, 1985) ist die bekannteste, LOOM (R. MacGregor, 1987) eine aktuelle Variante
- Konzepte werden in einer Taxonomie eingeordnet (Genus)
- Konzepte werden durch ihre Rollen (Slots, Attribute,..) definiert (Differentiae)
- Eine Rolle hat einen Namen sowie eine Werte- (*value restriction*) und Anzahlrestriktion (*number restriction*)
- Ein Subkonzept kann ein Konzept auf verschiedene Weise spezialisieren
  - Einführung neuer Rollen
  - Restringierung der Werterestriktion
  - Restringierung der Anzahlrestriktion

Beispiel

```
Geige isa Instrument
     baujahr: <1600-2004>

Stradivari isa Geige
     baujahr: <vor 1900>

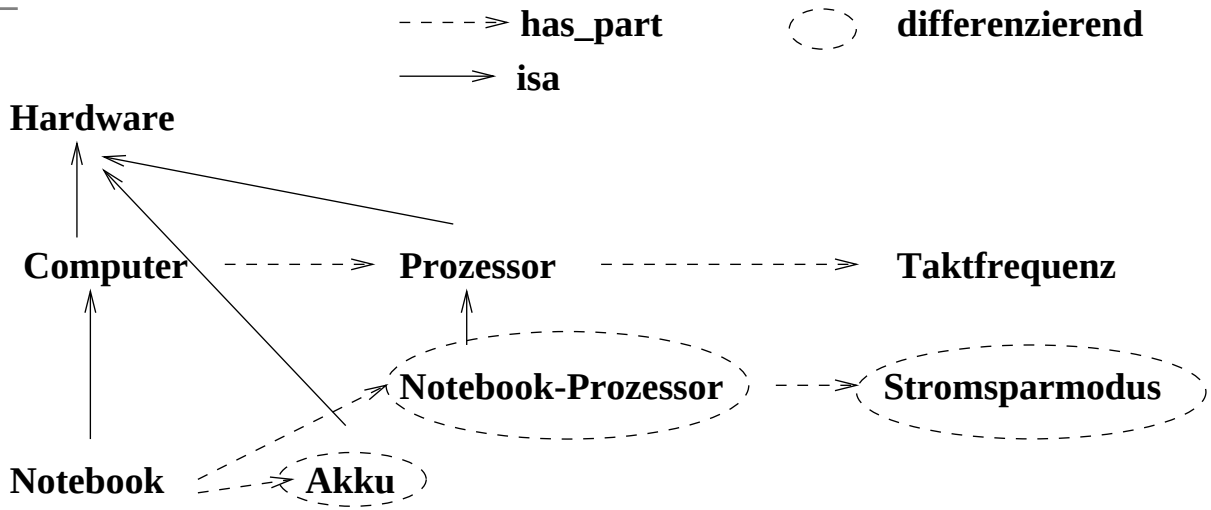
Musiker isa Kuenstler
     spielt: <mind. 1><Instrument>

Geiger isa Musiker
     spielt: <Geige>

Nobelgeiger isa Geiger
     spielt: <Stradivari>

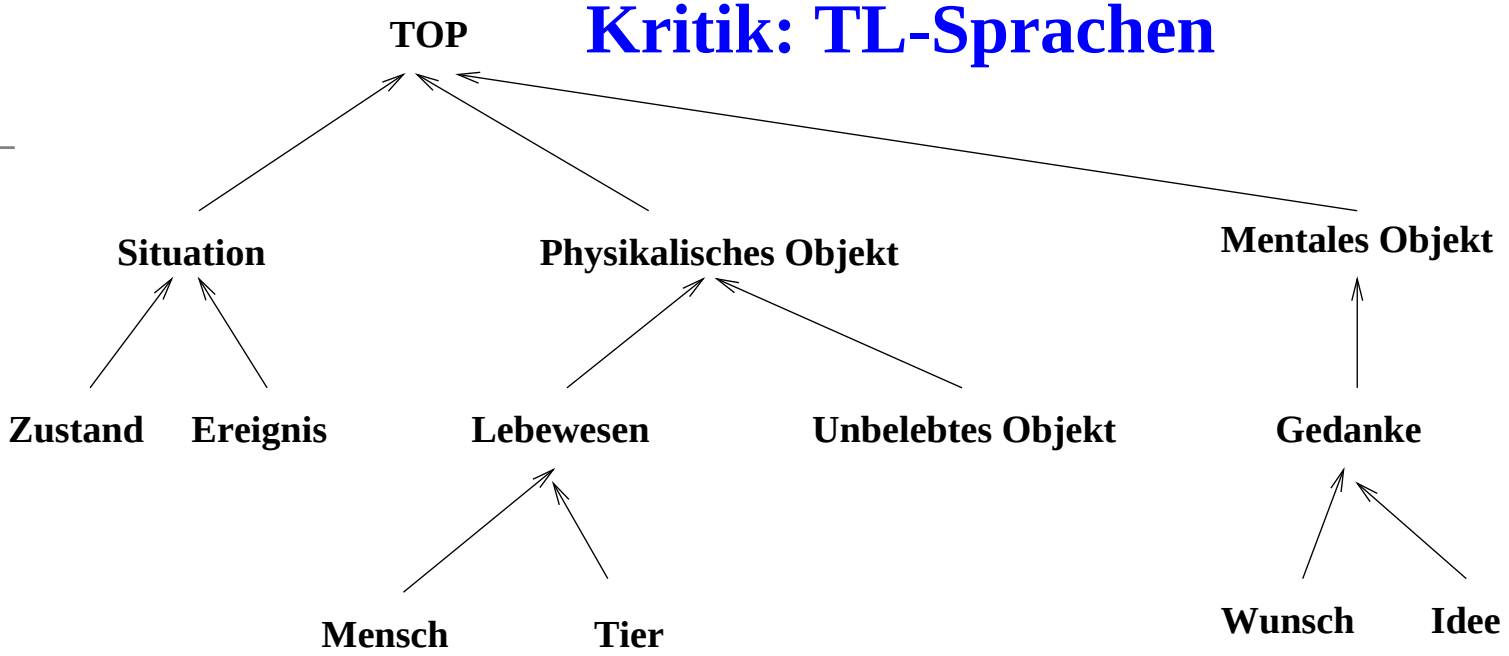
Allroundmusiker isa Musiker
     spielt: <mind. 2><Instrument>
```

Vererbung



- Vererbung: Weitergabe von Eigenschaften entlang der ISA-Kante
  - z.B. *Notebook* erbt die Rolle *Prozessor* und spezialisiert sie zu *Notebook-Prozessor*
  - *Notebook-Prozessor* ist spezieller als *Prozessor*, da er eine Rolle *Stromsparmodus* hat
  - *Notebook-Prozessor* erbt alle sonstigen Eigenschaften von *Prozessor*
- TL-Systeme checken (mit Einschränkungen) die Konsistenz von Wissensbasen (erleichtert Engineering)
- TL-Systeme klassifizieren neue Objekt aufgrund von Eigenschaften
  - Beispiel von oben: Ein Musiker, von dem bekannt ist, dass er 5 Instrumente spielt, wird automatisch als Allroundmusiker erkannt
  - Aber: diffizile Abhängigkeiten (z.B. closed-world assumption oder nicht)

Kritik: TL-Sprachen



- Idee terminologischer Modellierung: genus - differentiae
  - es gibt unterscheidende Merkmale zwischen Ober- und Unterbegriff
  - die Unterbegriffe eines Oberbegriffs unterscheiden sich hinsichtlich ihrer Merkmale
- funktioniert gut in technischen Domänen, nicht mehr so gut in Alltagsdomänen
  - z.B. abstrakte Begriffe wie Wut, Freude
- Problem der (Modellierungs-) Granularität
  - Computerkonstrukteur: Computer hat Prozessor? Nie!
  - Computerkonstrukteur: Computer hat Platine hat Prozessor. Eher!
- Frage immer: Modellierungszweck, gibt es eine neutrale Modellierung?

Kritik Wortnetze, Fazit

- auch hier: Frage der Granularität, der Modellierungsphilosophie, ...
- ist Beschränkung auf fixes Set an Sinnrelationen haltbar-ausreichend?
- bisherige Wortnetze unvollständig (z.B. keine Meronyme von Computer in GermaNet)
- Weltwissen statt Wortwissen?
- Weltwissen: CyC
- Wittgenstein's Sprachspiel: Bedeutung und Kontext
- es bleibt viel zu tun

Selektionsrestriktionen

- Verben haben Aktanten (DG), Verben haben Komplemente, Verben haben Argumente, Verben haben semantische Rollen
- Aktanten etc. eines Verbs sind nicht beliebig, d.h. bestimmte Objekte sind als Aktanten auszuschliessen
- wenn man die Klasse (den Typ) der Argumente, die erlaubt sind, (oft via einer Taxonomie) beschränken kann, dann spricht man von Selektionsrestriktionen des Verbs
- ‘geben’: agens = <Lebewesen>, rezipient = <Lebewesen>, theme =<???>
- Rollen sind *typisiert* (Füller müssen Typrestriktion entsprechen)
- Semantische Rollen(Kasusrollen, Tiefenkasus, Thetarollen)
  - aber nicht: grammatische Funktionen wie Subjekt, Objekt
  - Agent, Patient, Rezipient, Theme, Experiencer, Instrument ...
  - Ausgangspunkt Fillmore(1968): The case for case
  - guter Überblick in Allen(1995): Natural language understanding

Semantische Rollen

- Agent (Agens): intentionaler Verursacher
- Theme, Patient: das betroffene Ding (Lebewesen)
- Experiencer: die Person einer Wahrnehmnung o. ä.
- Instrument: Mittel/Kraft einer Aktion
- Rezipient: Endbesitzer, Empfänger
- Path: Route, Wegstrecke
- Source: Ursprung, Goal, Destination: Ziel
- u.v.m.: Giver, Sender, Sayer als spezifische Rollen
- Beispiele:
  - Tom (agent) zerstört den Rechner (theme)
  - Tom (theme) ist wütend
  - Tom (experiencer) genoss die Konferenz (theme)
  - Tom (agent) verkauft Thea (rezipient) seine E-Gitarre (theme)

Selektionsrestriktionen: wozu?

... verringern Ambiguitäten: Zusammenspiel verschiedener Wissensquellen bei der semantischen Interpretation

- Beispiel: ’das Buch gab sie dem Kind’
- ’Buch’ kann nicht Agens von ’geben’ sein (semantisch gesehen)
- ’Kind’ kann wegen Dativ nur Rezipient sein
- ’sie’ ist wegen Personalpronomen erlaubter Filler von Agens-Slot
- Slang: Filler (Wert) und Slot (Attribut)
- Abbildung: gram. Fkt. - Kasusrollen
  - teilweise kanonisch, aber auch verbspezifisch
  - Agens im Aktivsatz: Nominativ
  - Rezipient: Dativ

Selektionsrestriktionen: Probleme

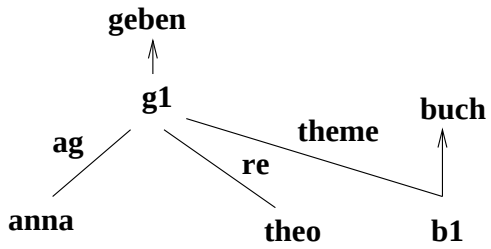
- Problem 1: oft gibt es keine natürliche Klasse, die als Selektionsrestriktion geeignet ist
  - essen: theme=<food> ??
  - das Kind isst Grashalme (theme=<alles was sich kauen und schlucken lässt>?)
  - Frage: sollte Typ eine Art 'normales Füllerobjekt' denotieren?
- Problem 2: oft werden Selektionsrestriktionen verletzt (figurative Sprache)

- 'Das Schnitzel an Tisch 5 möchte zahlen' (Metonymie)
- 'Bern hat beschlossen ..' (Metonymie)
- 'Sie ist eine Löwin' (Metapher)

- Behauptung: die Verletzung einer Selektionsrestriktion dient als Trigger (Auslöser) für eine figurative Interpretation
- Konversationsmaximenverletzung (Grice): 'Sei relevant'

Darstellungsvarianten

- ’Anna gibt Theo das Buch’
- Kasusframe: Slot-Filler (geben: agent=anna, ...)
- Prädikatenlogik: Thetarollen als Prädikate
  - geben(g1) & agent(g1,anna) & rezipient(g1,theo) & buch(b1) & theme(g1,b1)
  - basiert auf: Davidson, D. (1967): The logical form of action sentences
- graphische Form: semantisches Netz



Defaults und Prototypen

- 'Default' meint ganz allgemein: vorgegebener Wert, auch Voreinstellung oder normale (Werte-) Belegung, üblicher Wert
- Defaultwerte können überschrieben werden
- Beispiel: Theo kommt immer zu spät (Defaultfall) vs. er kommt pünktlich
- Beispiel: Jeder Vogel kann fliegen (Defaultfall) vs. Pinguin
- Prototypentheorie (Rosch 1973)

Das generative Lexikon

James Pustejovsky (1996). The Generative Lexicon. MIT Press, MA.

- Ziel: Modellierung polysemer Wörter und für die kreative Verwendung von Wörtern
- Spezifikation eines Kernlexikons (core lexicon)
- Einsatz generativer Mechanismen (generative devices) zur Erfassung abgeleiteter Wortbedeutungen
- versucht Enummerationsansatz zu ersetzen (Lexikoneintrag pro Bedeutung):
  - Schule wird gestrichen (Bedeutung 'Gebäude')
  - Schule war langweilig (Bedeutung 'Summe von Ereignissen')
  - Schule war reformiert (Bedeutung 'Bildungsinstitut')

$\rightarrow$ diese Bedeutungen sollen aus Kernbedeutung abgeleitet werden

$\rightarrow$ ähnlicher Ansatz: Bierwisch

Semantische Struktur von Wörtern

- Argumentstruktur: Anzahl und Typ der Argumente des Wortes
- Ereignisstruktur: Ereignistyp und interne Struktur
- Qualia-Struktur: Auflistung der Prädikationsweisen des Wortes
- Lexikalische Vererbungsstruktur: Bezug zu anderen Teilen des Lexikons (wird nicht betrachtet)

Argumentstruktur

- true arguments: syntaktische Realisierung ist obligatorisch
  She gives him a book (alle obligatorisch)
- default arguments: syntaktische Realisierung ist fakultativ
  she reads (das 'was' darf fehlen)
- shadow arguments: syntaktische Realisierung nur unter bestimmten Bedingungen pragmatisch valide: Mary buttered her toast with an expensive butter (with nur wegen expensive möglich)
- adjuncts: nicht speziell an das Wort gebunden (z.B. Ortsangaben)

Ereignisstruktur

Aktionsarten von Verben: ontologischer Gesichtspunkt (im Anschluss an Vendler (1967))

- accomplishments: durative Ereignisse mit Abschluss (Apfel essen, Haus bauen)
- achievements: punktuelle Ereignisse (Gipfel erreichen, etwas zerreissen)
- activities/ processes (wandern, regnen, singen)
- states (liegen, stehen, wohnen ..)
- Ermittlung anhand sprachlicher Tests, Überführbarkeit ineinander
- Pustejovski fasst accomplishments und achievements zu transitions (Übergänge) zusammen

Qualia-Struktur

- Const(itutive): Beziehung zwischen Objekt und seinen Teilen (part-of?); z.B. Material, Gewicht, Komponenten
- Form(al): (unterscheidende) Beziehung zu anderen Objekten (isa-Relation?); Farbe, Dimensionalität, Orientierung, Grösse
- Telic: Sinn und Zweck, Funktion (purpose-Relation?)
  - Zweck, den ein Agent mit einer Aktion bzgl. des Objekts verfolgt
  - die “built-in” Funktion, die bestimmte Aktivitäten spezifiziert
- Agentive: zur Entstehung gehörende Teile (cause-Relation?)
- ```
  novel(x)
       Argstr:    Arg1 = x:information
                  Arg2 = y:physical object
       Qualia
  Const:       hold(Y,x)
  Telic:      read(T,z,x)
  Agentive:  artifact(x), write(T,z,x)
  ```
  - Novelle ist ein entweder ein Informations- oder ein physikalisches Objekt, wobei das Physikalische Objekt die Informationen enthält (Const). Der Zweck ist gegeben als Transition (das T in read) vom Typ ’lesen’ und unter dem Agentive Aspekt gilt: es wird von jmd. geschrieben
  ```

Zwei generative Mechanismen (device)

- type coercion (Typerzwingung): Umwandlung eines Argumenttyps in einen anderen:

  'A semantic operation that converts an argument to the type that is expected by a function, where it would otherwise result in a type error' (Pustejovsky, 1996)

  Bsp.: 'Sie begann eine Novelle'
  $\rightarrow$ nur Ereignisse können begonnen werden
  $\rightarrow$ die Telic- oder Agentive-Struktur liefert $read$ bzw. $write$ als mögliche Transitionen (Ereignisse)
  $\rightarrow$ je nach Kontext wird "Lesen oder Schreiben einer Novelle" verstanden

- co-composition: unter bestimmten Umständen (Ko-Spezifikation) werden Teile von Qualia-Strukturen zusammengefasst

  Bsp.: 'Er kocht Kartoffeln' vs. 'Er kocht Paella'
  $\rightarrow$ Zustandsveränderung (Kartoffel) vs. Zustanderzeugung (Paella)

type coercion: 'Sie begann eine Novelle'

- Argumentstruktur von 'beginnen' fordert:
  arg1 = <human> arg2 = <transition>
  - arg1 wird korrekt belegt ('Sie' erfüllt <human>, z.B. kontextuell)
  - arg2 wird verletzt, da 'Novelle' kein Ereignis ist
- die Qualiastruktur von 'Novelle' enthält 2 Bestandteile, die die Forderung von 'beginnen' erfüllen: read und write
- 'Novelle' kann als 'Schreiben bzw. Lesen einer Novelle' uminterpretiert werden

Kritik: Theorie ist z.B. übergenerell

Pustejovsky's Theorie ist einflussreich ist, hat aber Schwachstellen

- John began the film (watching? casting?)
- John began the door (opening?, closing?, walking through?)
- John began the nails (hammering in?)
- Sind solche Alternationen stärker konventionalisiert?

Verspoor: Empirie

Korpusuntersuchungen zu ’begin’ und ’finish’ anhand geschriebener und gesprochener Sprache

- Lancaster-Oslo/Bergen Korpus (LOB): British English, geschriebene Sprache, ca. 1 Mio Wörter
- British National Korpus (BNC): British English, gesprochene Sprache (ca. 90 Mio), geschriebene Sprache (ca. 10 Mio)
- Resultate
  - sehr wenig Metonymien mit ’begin’ (3 LOB, 164 BNC), deutlich mehr für ’finish’
- Verspoor, C. (1997): Contextually Dependent Lexical Semantics. PhD thesis, University of Edinburgh
  - Telic-Metonymien treten nur in 20 Klassen von Nomen auf: food,story, written object,game
  - Agentive-Metonymien sind v.a. für Artifakte gut möglich (als Entstehen)

TEIL II: Wortvorkommen

Rep.: Types and Tokens

- Token: in einem Text vorkommende Wortformen
- Types: in einem Text vorkommende unterschiedliche Wortformen
- Bsp.: ’Die Frau sah das Mädchen, aber das Mädchen hat sie nicht gesehen.’
- hat 10 Types: ’das’ und ’Mädchen’ werden nur einmal gezählt
- hat 12 Token
- oder 12 Types, wenn man die Satzzeichen eigens zählt (14 Token)
- ’sah’ und ’gesehen’ können auch als zwei Instanzen eines Lexems aufgefasst werden

Zipfs Gesetz

- Die Häufigkeit des Auftretens eines Tokens ist invers proportional zu seiner Position in der Häufigkeitsliste
- $freq \approx \frac{1}{rank}$ oder: $freq * rank = constant$
- so dass z.B. ein Wort an Position 50 etwa 3 mal so häufig auftritt wie eines an Position 150
- Gesetzmässigkeit stimmt nicht ganz am Anfang und am Ende der Liste
- wichtige Aussage: die meisten Wörter sind rar (sparse data problem)
- im Brown-Corpus (nach Smith S.79): 1 Mio Token (ca. 50.000 Types), davon machen 6 Types (the, of, and, to, a, in) 205.961 Token aus

Zipfs Gesetz fortg.

rank	$*$	frequency	constant
35	very	836	29.260
45	see	674	30.330
55	which	563	30.965
65	get	469	30.485
75	out	422	31.650

Anmerkung: Zweites Gesetz von Zipf: Die Häufigkeit des Auftretens eines Wortes ist umgekehrt proportional zu seiner Länge.
Effizienz $\leftrightarrow$ Expressivität

Zipfs Gesetz fortg.

nach Meier, H.: Deutsche Sprachstatistik. Hildesheim: Olms. 1964

rank	type	freq	Obs
1	die	349.553	
2	der	342.522	
3	und	320.072	1. Konjunktion
4	in	188.078	1. Präposition
5	zu	172.625	
6	den	138.664	
7	das	124.232	
8	nicht	114.518	1. Adverb
9	von	113.201	
10	sie	102.212	1. Personalpronomen
11	ist	96.970	1. Hilfsverb
12	des	96.190	
13	sich	92.945	
14	mit	91.552	
15	dem	89.109	
...			
90	Zeit	14.529	1. Substantiv
...			
127	machen	8.929	1. Vollverb

Obs: Die häufigsten Wörter sind Funktionswörter, vor allem Determiner und Präpositionen.

Kollokationen

- Kollokation: weite und enge Definitionen, z.B. Mehrwortphrase mit spezieller Bedeutung, die nicht kompostional ermittelbar ist
- idiomatische Ausdrücke, Mehrwortfachterme, Komposita (im Englischen, ...), Verb-Präposition-Kombinationen
- kompositional interpretierbar: ’Löffel abgeben’: Löffelabgebaktion
- nicht kompositional interpretierbar: ’Löffel abgeben’: sterben
- nicht substituierbar: ’white wine’ vs. ’yellow wine’
- nicht modifizierbar: ’den goldenen Löffel abgeben’ (stilvoll sterben”)
- weit: auch Verben mit (semantisch leerer) Präposition wie z.B. ’warten auf (jmd.)’
- Übersetzung in andere Sprache, wenn keine Wort-zu-Wortübersetzung möglich, dann Kollokation

Kollokationsfindung

- gemeinsame Auftretenshäufigkeit $freq(w1\ vor\ w2)$
- Idee: Kollokation falls sehr hoher Wert
- Problem: häufige Wörter sind meist Funktionswörter
- Lösung I: Normalisierung (z.B.
  $score = freq(w1\ vor\ w2)^2 / freq(w1) * freq(w2)$
  - Normalisierung heisst hier: Vorkommenshäufigkeit der einzelnen Wörter berücksichtigen
  - wie funktioniert die Formel: wenn zwei Wörter nur zusammen auftreten, dann ist der Score 1
  - aber: zwei sehr seltene Wörter, die zufällig im Text auftreten, sind nicht gleich eine Kollokation, daher Frequenzminimum für gemeinsames Auftreten setzen
- Lösung II: Einsatz von Tagger (nur Inhaltswörter betrachten)

Beispielprogramm in Perl

```
$w="[a-zA-ZöäßüÜÄÖ]";
$lastword=’’;
$wort{’’}=1;

while(<>){ # Einlesen der Daten, Hash(es) aufbauen, zaehlen
 while(/($w+) /g){
  $bigram{"$lastword $1"}++;
  $wort{"$1"}++;
  $lastword=$1;
 };
};

foreach (keys %bigram){ # score = freq(w1 w2)^2/freq(w1)*freq(w2)
 /($w+) ($w+)/;
 $zaehler = $bigram{"$1 $2"} * $bigram{"$1 $2"};
 $nenner = $wort{$1} * $wort{$2};
 next if $nenner eq 0;
 $score{"$1 $2"} = $zaehler/$nenner;
};
```

Kollokationen im Stechlin

```
frische Luft  = 0.25| frische Luft  = 3| frische=3 | Luft=12
Baron Beetz  = 0.2580| Baron Beetz  = 8| Baron=31 | Beetz=8
Friedrich Wilhelms = 0.2976| Friedrich Wilhelms = 5|Friedrich=14 |Wilhelms=6
en beten  = 0.333| en beten  = 6| en=18 | beten=6
Ick weet  = 0.333| Ick weet  = 4| Ick=8 | weet=6
serr gutte  = 0.333| serr gutte  = 3| serr=9 | gutte=3
Baruch Hirschfeld  = 0.3629| Baruch Hirschfeld  = 7| Baruch=15 | Hirschfeld=9
Rolf Krake  = 0.375| Rolf Krake  = 6| Rolf=16 | Krake=6
junges Ding  = 0.375| junges Ding  = 3| junges=3 | Ding=8
Mamsell Pritzbur  = 0.444| Mamsell Pritzbur  = 4| Mamsell=9 | Pritzbur=4
Christum erkennen  = 1| Christum erkennen  = 3| Christum=3 | erkennen=3
Cremmer Damm  = 1| Cremmer Damm  = 7| Cremmer=7 | Damm=7
```

Keine Kollokationen

```
ihn ganz  = 0.00015475351204498 2| ihn ganz  = 3| ihn=187 | ganz=311
man am  = 0.00015479576633579 1| man am  = 4| man=483 | am=214
und daneben  = 0.00015487334801762 1| und daneben  = 3| und=3632 | daneben=16
und stieg  = 0.00015487334801762 1| und stieg  = 3| und=3632 | stieg=16
und außerdem  = 0.0001548733480 17| und außerdem  = 3| und=3632 | außerdem=16
es daß  = 0.00015499526688007 3| es daß  = 11| es=1143 | daß=683
mehr Und  = 0.00015499537166598 5| mehr Und  = 6| mehr=205 | Und=1133
ich dann  = 0.00015551109042965 72| ich dann  = 8| ich=1344 | dann=307
man auf  = 0.00015551212161503 35| man auf  = 7| man=483 | auf=654
auch gern  = 0.00015551831160769 71| auch gern  = 3| auch=1074 | gern=54
war nur  = 0.00015551951579110 73| war nur  = 6| war=789 | nur=294
der wo  = 0.0001555342514392 96| der wo  = 7| der=2022 | wo=156
mir eine  = 0.00015554082055532 53| mir eine  = 6| mir=381 | eine=608
ich hätte  = 0.00015557308970070 09| ich hätte  = 3| ich=1344 | hätte=43
macht die  = 0.00015557859889976 15| macht die  = 4| macht=41 | die=2505
```

Wozu Kollokationen?

- Indexierung beim IR (Index: Zugriffsschlüssel für Datenabruf)
- Sprachlernen: idiomatische Ausdrücke, Redewendungen
- Übernahme in Wörterbuch, Fachwörterbuch (Terminologie)
- Übersetzung (oft nicht Wort-zu-Wort übersetzbar):
  - feeling blue - *sich blau fühlen (statt: traurig sein)
  - on the other hand - *in/auf der anderen Hand (statt: auf der anderen Seite)

Konkordanzen

- ist definiert als: Auflistung eines Items (Wort, Phrase) inklusive seines Kontextes
- Grundlage für (lexikographische) Korpusarbeit verschiedenster Prägung (z.B. Subkategorisierung)
- KWIC (Konkordanzprogramm): Key Word in Context bedeutet spezielle Ausgabe (Lesbarkeit)
- Konkordanzprogramme als Grundlage von COBUILD Dictionary, des ersten Wörterbuchs dessen Beispielsätze komplett aus real-existierenden Texten stammten.

Lexikon: in der CL

- das Lexikon wird in CL zur automatischen Sprachverarbeitung verwendet
  - der Aufbau der Lexikoneinträge ist meist theorien- und implementationsabhängig
  - oft nur Teilprobleme gelöst/erfasst (spezieller Anwendungskontext)
  - oft nur prototypisch (Skalierbarkeit)
  - der Lexikonaufbau kann auch Ziel eines (CL-) Projektes sein (z.B. Celex)
  - Ziel ist es dann, Ressourcen für Anwendungen etc. zu schaffen
  - der Aufbau der Lexikoneinträge ist auch hier theorien- und implementationsabhängig

... und sonst

- Lexika werden auch ausserhalb der CL (und das schon seit langer Zeit) gebraucht
  - das traditionelle Lexikon ist ein Nachschlagewerk für Menschen
  - es hat sich die Wissenschaft (Lehre) der Lexikonschreibung (Lexikographie) entwickelt
  - mit Einzug des Computers sind maschinenlesbare Lexika entstanden
  - diese sind aber eher nicht (zumindest nicht direkt) maschinenverwertbar (CL-verwertbar)
  - aber: Verfahren der Computerlinguistik werden mittlerweile auch von Lexikologen verwendet
  - womit dieser Bereich zu einem wichtigen Berufsfeld für CLInnen ist

Grundlegende Unterscheidungen

- **Enzyklopädie** [zu griech. kyklos „Kreis“ und paideía „Bildung“], 1. nach Hippias von Elis, einem Sophisten des 5.Jh. v.Chr., Begriff für die universale Bildung, später allg. die Alltagsbildung, ... in der Neuzeit dann seit dem 17./18.Jh., zunächst unter Einfluß der -> Enzyklopädisten, Begriff für die Gesamtheit des menschl. Wissens. - 2. die Darstellung der Bildungsinhalte und Wissensgebiete bzw. -bereiche sowie einzelner -gegenstände.
- **Lexikon** [zu griech. lexikón (biblíon) „das Wort betreffendes (Buch), Wörterbuch“], nach Stichwörtern geordnetes Nachschlagewerk, das entweder ein oder mehrere Sach- und Wissensgebiete (-> Enzyklopädie) oder den Wortschatz einer oder mehrerer Sprachen, von Fach-, Sonder-, Gruppensprachen usw. auflistet (-> Wörterbuch).
- **Wörterbuch**, Nachschlagewerk, das den Wortschatz einer Sprache nach bestimmten Gesichtspunkten auswählt, anordnet und erklärt; das W. gibt Sprachinformation, während das Lexikon Sachinformationen bietet.

Lexikographie, Lexikologie

- Die **Lexikologie** analysiert den Inhalt eines Lexikons. Sie untersucht und beschreibt den Wortschatz einer Sprache, in Hinblick auf die Bedeutungsstruktur und die Zusammenhänge zwischen den einzelnen Wörtern.
- Die **Lexikographie** dagegen beschäftigt sich mit der formalen Darstellung des Inhalts eines Lexikons. Sie berücksichtigt dabei einerseits die theoretischen Erkenntnisse der Lexikologie andererseits die Anforderungen der Benutzer. Sie umfasst auch den Vorgang und die Methode der Anfertigung von Wörterbüchern, sowie die Art der Organisation und (in CL) Abspeicherung.

Wörterbuchtypologie

- Anzahl Sprachen
  - einsprachig, zweisprachig, mehrsprachig
  - unidirektional, bidirektional
- Art der Finanzierung
  - akademische Wörterbücher (Middle English Dictionary)
  - kommerzielle Wörterbücher (Webster)
- Benutzungssituation
  - Es gibt Schüler-Wörterbücher für verschiedene Altersklassen. Sie unterscheiden sich im Umfang und in der Aufmachung (Bebilderung). Problem: Wie ermittelt man den Wortschatz der jeweiligen Altersklasse? Bsp.: wissenschaftl. Lexikon, Schülerlexikon, allgemeines Lexikon, Sprachlernlexikon, Konversationslexikon (’Urlaubslexikon’)

Wörterbuchtypologie fortg.

- Grösse: Maximalschätzungen für den englischen Wortschatz belaufen sich auf 4 Mio Wörter (700.000 in den Merriam-Webster Dateien, 1 Mio wissenschaftl. Wörter, dazu Dialektwörter, Slang, Neologismen, Handels- und Ortsnamen). Probleme: Es gibt Millionen chemischer Substanzen mit je eigenen Namen.
  - English unabridged dictionary: 400.000 - 600.000 Wörter (allgemeiner Gebrauch)
  - College dictionary: 130.000 - 160.000 Wörter
  - Desk dictionary: 60.000 - 100.000 Wörter
  - Pocket dictionary: 40.000 - 60.000 Wörter

Wörterbuchtypologie fortg.

- Bereich (Grundgesamtheit)
  - Es gibt bereichsspezifische, fachsprachliche Wörterbücher vor allem für Jura, Medizin, Biologie, Elektronik und Architektur. Aber auch: wirtschaftswissenschaftliches Lexikon, regionalsprachliches Lexikon (z.B. Duden: Wie sagt man in der Schweiz?), Fremdwörterbuch
  - Die Qualität dieser Wörterbücher ist sehr unterschiedlich
  - Der Anteil technischen Vokabulars in einem allgemeinen Wörterbuch nimmt stark zu (ca. 40
- Sprachaspekte (inhaltliche Tiefe): Etymologie, Aussprache, Orthographie, Gebrauch, Synonyme, Slang, Dialekt, Belege

Wörterbuchtypologie fortg.

- Zeitausschnitt: synchron vs. diachron (Problem: die Erstellung eines Wörterbuchs dauert of mehrere Jahrzehnte.) Bsp.: etymologisches Wörterbuch, Wörterbuch der Jugendsprache, 'Kleines Lexikon untergegangener Wörter'
- Linguistischer Ansatz: präskriptiv vs. deskriptiv
- Zugriffsarten
  - Wichtigstes Ordnungskriterium: alphabetische Ordnung
  - inhaltliche Ordnung (plus alphabetischer Index); z.B. Duden Bildwörterbuch
  - Sortierung nach Wortlänge; z.B. Kreuzworträtsellexikon
  - Rückläufig alphabetisch
  - Alter (Jahr des ersten Auftretens)
  - Zeichenaufbau (Anzahl der Radikale) im Chinesischen
- semasiologische (geordnet nach Wortfeldern) vs. onomasiologische (geordnet nach Sach- und Begriffsgruppen) Wörterbücher

Probleme der Grössenbestimmung

Grösse ist eines der wichtigsten Verkaufsargumente. Die Angaben sind deshalb mit Vorsicht zu interpretieren.

- Traditionelles Zählsystem: Jedes Lemma (engl. head word) ist ein Eintrag
- Amerikanisches Zählsystem: Jedes Wort und jede Phrase, die explizit oder implizit definiert ist und identifiziert werden kann (z.B. durch Fettdruck), ist ein Eintrag. Genauer:
  - Das Lemma ist ein Eintrag
  - Jede weitere Wortart zu dem Lemma ist ein Eintrag
  - Flektierte Formen, die aufgeführt werden, gelten als Eintrag
  - Hinzugefügte Derivationen (engl. run-ons) gelten als Eintrag
  - Hervorgehobene Redewendungen innerhalb eines Artikels gelten als Eintrag

Beispiel

Beispiel (leicht gekürzt) zählt als 5 Einträge:

parachute
n. An apparatus of lightweight fabric that when unfurled assumes the shape of a large umbrella and acts to retard the speed of a body moving or descending through air. v. **chuted, chuting** v.t. 1. to land (troops, materiel, etc.) by means of parachutes. v.i. 2. to descend by parachute –**parachutist.**

In amerikanischen College Dictionaries sind weniger als die Hälfte aller Einträge Lemmas.

Thesaurus

- In der Regel enthalten Thesauruseinträge keine syntaktischen und morphologischen Angaben, sondern nur einige semantische Relationen
- In der Makrostruktur werden die Elemente von Thesauri nach semantischen Kriterien geordnet; es handelt sich also um onomasiologische Lexika, während Lexika im allgemeinen alphabetisch geordnet sind. Für die Benutzung wird ein Thesaurus jedoch durch ein Register (semasiologisch) erschlossen
- Ein Thesaurus enthält meistens Substantive (evtl. Adjektive und Verben). Funktionswörter sind selten
- Thesauri informieren über die Relation eines Deskriptors zu einem anderen, nicht über dessen Bedeutung

Kodierungsfragen I

Herstellung und Nutzung von Lexika heutzutage (oft) elektronisch

- Im Computer alle Daten als Bitmuster (= Zahlen) kodiert: welchen Zeichensatz (ASCII, ISO-8859-X, Unicode etc.) verwenden?
  - **ASCII**: am weitesten verbreitete Kodierung; Mängel:nur für die lateinische Schrift geeignet, kodiert nur die 26 im Englischen gebräuchlichen Buchstaben
  - **Unicode** ist der Versuch, möglichst alle Schriften möglichst aller Sprachen in einer Kodierung unterzubrigen:
  - zusammengesetzte Zeichen: Diakritika auf Basiszeichen (z.B. a + ¨ = ä; aus zwei oder mehr Teilzeichen zusammengesetzte Zeichen in Hangul (Koreanisch))
  - Grafikzeichen (Rahmen etc.), diverse Sonderzeichen (z.B. Währungssymbole)
  - 16-Bit-Kodierung; daher potentiell 65536 Codepunkte; im Moment umfaßt Unicode ca. 38000 Zeichen

Kodierungsfragen II

- welches Format (SGML/XML-basierte vs. proprietäre Auszeichnung)
  - SGML: Standard Generalized Markup Language; Metasprache (!) zur Definition von Dokumentbeschreibungssprachen; Beispiele: HTML, TEI
  - XML, Extensible Markup Language (echte Teilmenge von SGML)
  - TEI: Text Encoding Initiative; SGML-basierte Dokumentbeschreibungssprache für die Auszeichnung von Text; hat bereits Auszeichnungsformate für verschiedenste Textsorten, z.B. Prosa, Lyrik, Drama, gespr. Sprache, Wörterbücher (siehe nächste Woche)

Vorteile vom SGML/XML

- Vorteile von SGML/XML:
  - SGML ist (XML weitgehend) standardisiert ==> leichter Datenaustausch
  - große Auswahl an bereits existierender Verarbeitungssoftware
  - mit Hilfe von SGML (XML) können Beschreibungssprachen für beliebige Phänomene erzeugt werden

Zugang zum Sprachgebrauch

- Ziel: fortlaufende Aktualisierung des Wissens über den aktuellen Sprachgebrauch (z.B. für elektronisch verfügbare Wörterbücher, Bedeutungshotlines, sonstige sprachbezogene Dienste)
- Grund: fortlaufender Bedeutungswandel, Wortneuschöpfungen
- aber wie: Introspektion versus empirische Exploration? Individuelle Sprachkompetenz versus irgendwie fixierbarer Sprachstandard?
- SprecherInnen “sind nicht gleichzeitig in allen Dialektgebieten zu Hause, beherrschen nicht alle Fach- oder Gruppensprachen, haben unterschiedliche Sprachbiografien oder unterschiedliche Bewertungen des Fremdwortgebrauchs in der Muttersprache” (http://www.ids-mannheim.de/elexiko/Korpusbasiertheit.html)
- Problem: individuelle sprachliche Besonderheiten von wirklich Üblichem zu unterscheiden
- was tun? Korpuslinguistik!

Fragen der Lexikographie an die KL

(siehe: http://www.ids-mannheim.de/elexiko/Korpusbasiertheit.html)

- Kommt ein Wort überhaupt vor und wenn ja wie häufig?
- In welchen Situationen kann man dieses Wort gebrauchen?
- Wann ist dieses Wort zum ersten Mal belegt?
- Welche Bedeutungen und Kontexte finde ich als Nicht-Muttersprachler(in) für dieses Wort?
- Gibt es historische Perioden, in denen dieses Wort besonders häufig verwendet wurde?
- Gibt es ein mir bekanntes Wort überhaupt noch im alltäglichen Gebrauch?
- Kommt dieses Wort nur in bestimmten Textsorten vor?
- Wird es nur von einer bestimmten Sprechergruppe verwendet?
- ... und viel mehr! (siehe Hans-Martin Lehmanns Vortrag)