Lexikologie, Lexikographie und Lexikonstrukturen

Morphologieanalyse und Lexikonaufbau (11. Vorlesung)

Übersicht

Motivation

1. Wie kann die Wörterbucherstellung von der Computerlinguistik profitieren?

1. Wie kann die Computerlinguistik (NLP) von der Information in Wörterbüchern profitieren?

Definitionen

Enzyklopädie - Lexikon - Wörterbuch

- **Enzyklopädie** [zu griech. kyklos ,,Kreis`` und paideía ,,Bildung``], 1. Seit 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.

Lexikologie - Lexikographie

- 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.

Wörterbuch-Typologie

- **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). Lexikon, Schülerlexikon, allg. Lexikon, Sprachlernlexikon, Konversationslexikon
- **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

- **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% technisch-wissenschaftl. Wortschatz in einem College Dictionary).
- **Sprachaspekte (inhaltliche Tiefe)**: Etymologie, Aussprache, Orthographie, Gebrauch, Synonyme, Slang, Dialekt, Belege
- **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 (z.B. Gebrauchsempfehlungen) vs. deskriptiv
- **Zugriffsarten**
  - Wichtigstes Ordnungskriterium: alphabetische Ordnung
  - inhaltliche Ordnung (plus alphabetischer Index); z.B. Duden Bildwörterbuch
  - Sortierung nach Wortlänge; z.B. Kreuzworträtsellexikon
  - Häufigkeit
  - Rückläufig alphabetisch
  - Alter (Jahr des ersten Auftretens)
  - Zeichenaufbau (Anzahl der Radikale) im Chinesischen
- **semasiologische** (geordnet n. Wortfeldern) vs. **onomasiologische** (geordnet n. Sach- & Begriffsgruppen) W.

Semasiologie

Teildisziplin bzw. Forschungsrichtung der Semantik, die sich mit der Bedeutung einzelner sprachlicher Ausdrücke, den Bedeutungsbeziehungen zwischen sprachlichen Ausdrücken (Wortfeld), sowie Problemen des Bedeutungswandels beschäftigt.

semasiologische Wörterbuchordnung:

```
regeln, Regel, Regelung, regiere
n, Regierung, ...
```

Onomasiologie

Teildisziplin bzw. Forschungsric htung der Semantik, die sich - ausgehend von Sachverhalten und Begriffen der realen Welt - mit der Erforschung der auf sie referierenden sprachlichen Ausdrücke (= Wörter) beschäftigt. Dabei werden auch Aspekte der geographischen Verteilung bestimmter Bezeichnungen (Wortatlas) berücksichtigt. Typischer Vertreter: **Thesaurus**

onomasiologische Wörterbuchordnung:

```
Regel, Gesetz, Heuristik, ...
```

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:

1. Das Lemma ist ein Eintrag.
2. Jede weitere Wortart zu dem Lemma ist ein Eintrag.
3. Flektierte Formen, die aufgeführt werden, gelten als Eintrag.
4. Hinzugefügte Derivationen (engl. *run-ons*) gelten als Eintrag.
5. Hervorgehobene Redewendungen innerhalb eines Artikels gelten als Eintrag.

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.

Probleme der alphabetischen Ordnung

Wörterbücher enthalten eine buchstaben-orientierte alphabetische Ordnung. Beispiel:

```
power
powerful
power of attorney
```

Vorteil: Der Benutzer muss nicht wissen, ob ein Kompositum zusammengeschrieben wird.

Besondere Probleme: Verbgefüge

```
have one's eye on
jmd. Bescheid geben
```

Wörterbucheintrag: Inhalt des Lexikons (nach Lorenz & Schreiber 1998, Uni Erlangen)

Hudsons „All-Inclusive Lexicon``

1. ***Phonologie***
   - phonologische Segmentstruktur bzw. mehrere solcher Strukturen, falls Allomorphie auch im Lexikon beschrieben wird
   - prosodische Information (soweit diese nicht regelhaft hergeleitet werden kann), also beispielsweise Betonung oder Tonalität.
2. ***Morphologie***
   - morphologische Segmentstruktur(en)
   - Allomorphieverhalten
   - morphosyntaktisches Verhalten (z.B. Flexion)
   - „verwandte`` Wörter (Ableitungen, Zusammensetzungen)
   - eventuell Klitisierung
3. ***Syntax***
   - allgemeine Wortklasse (z.B. „Verb``);
   - Unterklasse (z.B. „Hilfsverb``);
   - Valenz (obligatorische oder fakultative Ergänzungen)
4. ***Semantik***
   - Synonymie und Antonymie
   - Hyponymie und Hyperonymie
   - semantische Rollen abhängiger Ergänzungen
5. ***Kontext***
   - stilistische Restriktionen (z.B. „formell``/„vulgär``/„Slang``)
   - soziolinguistische Restrikionen
   - restrictions relating to larger social structure (Klassifikation des Sprechers)
   - Restriktionen bezüglich der Diskursstruktur

6. ***Orthographie***
   - normale Schreibung;
   - eventuelle abweichende Schreibungen
   - übliche Abkürzungen oder Ideographen
   - die Orthographie betreffende Unregelmäßigkeiten
7. ***Etymologie***
   - Die Sprache, der das Wort angehört (in einem mehrsprachigen Lexikon)
   - eventuell die Sprache, aus der das Wort entlehnt wurde
   - das Grundwort, von dem das beschriebene Wort abgeleitet wurde
   - der Zeitpunkt der Entlehnung
8. ***Verwendung***
   - Häufigkeit und Vertrautheit
   - Belege für die Verwendung des Wortes
   - häufige Kollokationen/idiomatische Verwendungen
   - Tabus.

TEI-Lexikon

1. ***Nouns***

· Entity nouns - *apple, book*, etc.;
· relational nouns - *speed*, *age*, *height*, *father*, etc.;
· abstract nouns - *courage*, *love*, *altruism*, etc.;
· mass nouns - *wine*, *sand*, etc.;
· proper names - *John*, *Europe*, *IBM*, etc.;
· complement taking properties: e.g. ``factive'' nouns like *story*, ``transitive'' nouns, etc.

2. ***Pronouns*** (*I,he*, *she*, *it*, etc.) and bound anaphors (*myself*, *himself*, *herself*, *each other*, etc.).
3. ***Verbs***

A wide variety of valency classes:
· Intransitive;
· transitive;
· ditransitive;
· clausal complement taking (raising and control);
· `small-clause' taking verbs including naked infinitives, etc.;
· an indication of variants of a basic valency class
· in a language like German, the nominal complements of a verb may appear in different cases (e.g. *helfen* vs. *sehen*)

4. ***Modals and auxiliaries***
5. ***Prepositions***

· subclasses of prepositions, e.g. ``case-making'' prepositions vs. semantical contenful;
· whether english particles are represented as a subset of prepositions.

6. ***Adjectives***

· Complement-taking properties: e.g. *proud of*
· semantic classes of adjectives distinguished
· the position in which an adjective can appear (pre-nominal, post-niminal, predicate position, etc.).

7. ***Determiners and other similar nominal modifiers*** (e.g. articles, quantifiers, demonstratives, etc.)
8. ***Multi-word lexical entries***
9. ***Inflected categories of noun, verb, and adjective***

Thesaurus

Unterschiede:

1. In der Regel enthalten Thesauruseinträge keine syntaktischen und morphologischen Angaben, sondern nur einige semantische Relationen.
2. 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.
3. Ein Thesaurus enthält meistens Substantive (evtl. Adjektive und Verben). Funktionswörter sind selten.
4. Thesauri informieren über die Relation eines Deskriptors zu einem anderen, nicht über dessen Bedeutung.

Wörterbuchorganisation: Form des Lexikons

nach Lorenz & Schreiber 1998, Seminar Uni Erlangen

Themenkomplexe

- Organisation und Zugriff (attributiert vs. tabellarisch)
- Auszeichnung (SGML-basierte vs. proprietäre Auszeichnung)
- Zeichensatzfragen (ASCII, ISO-8859-X, Unicode etc.)

Organisation und Zugriff

***Tabellarisch*:** jeder Eintrag besteht aus einer festgelegten Sequenz von Feldern, wobei jedes Feld eine festgelegte Bedeutung hat. Der Zugriff erfolgt über die Feldnummer.

Vorteil: sehr einfache Verarbeitung.

Nachteile: unübersichtlich; unstrukturiert; eventuell redundant.

Beispiel (CELEX):

```
7\Aas\6\M\1\Y\Y\Y\Aas\N\N\N\N\(Aas)[N]\N\N\N\N\S1/P1\Y
19\Abart\7\Z\1\Y\Y\Y\abart\V\N\N\N\(((ab)[V|.V],((Art)[N])[V])[V])[N]\N\N\N\N\S3/P3\N
```

***Attributiert*:** der Zugriff auf die verschiedenen Informationen im Eintrag erfolgt über Schlüssel (= Attributnamen), denen Werte zugeordnet sind; diese Werte können auch komplex (= geschachtelt) sein und verschiedene Datentypen enthalten.

Nachteil: aufwendigere Verarbeitung.

Vorteile: übersichtlich; weniger redundant; die Daten können in einer Art strukturiert werden, die den Charakter des zu beschreibenden Phänomens widerspiegelt.

Beispiel:

```
[Surface: [Lemma: "aas"],
 Form: [POS: Substantive,
        Combi: [SgDecl: Sg_e_s,
                PluralSx: Pl_e,
                PlDatSx: yes],
        Syn: [Gender: Neuter]]];
[Surface: [Lemma: "abart"],
 Form: [POS: Substantive,
        Combi: [SgDecl: Sg_0,
                PluralSx: Pl_en],
        Syn: [Gender: Feminine]]];
```

Auszeichnung ("Mark-Up")

SGML: Standard Generalized Markup Language; Metasprache zur Beschreibung von Dokumentbeschreibungssprachen; Beispiele: HTML, XML, PML, TEI.

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**

Mit SGML können eigene Dokumentbeschreibungssprachen erzeugt werden, die eventuell besser für das zu beschreibende Phänomen geeignet sind. Die Vorteile sind:

- SGML ist standardisiert ==> leichter Datenaustausch
- große Auswahl an bereits existierender Verarbeitungssoftware
- Mit Hilfe von SGML können Beschreibungssprachen für beliebige Phänomene erzeugt werden.

Wenn in best. Situationen auf SGML verzichtet werden muß, können auch proprietäre Formate zum Einsatz kommen.

- Nachteil: erschwerter Datenaustausch (da kein standardisiertes Format).
- Vorteil: das proprietäre Format ist eventuell optimal an die verwendete Software angepaßt.

Zeichensatzfragen

- Intern werden im Computer alle Daten als Bitmuster (= Zahlen) kodiert.
- Welche Zahl wird welchem Buchstaben zugeordnet?
- ASCII: am weitesten verbreitete Kodierung; hat mehrere Mängel:
    - nur für die lateinische Schrift geeignet,
    - kodiert nur die 26 im Englischen gebräuchlichen Buchstaben.
- Daher existieren diverse Standards für die Kodierung anderer Sprachen.

Unicode ist der Versuch, möglichst alle Schriften möglichst aller Sprachen in einer Kodierung unterzubrigen. Dabei sind Aspekte zu beachten wie:

Direktionalität: Schreibrichtung Links nach Rechts, Rechts nach Links, Oben nach Unten, wechselnd, etc.

Zusammengesetzte Zeichen: Diakritika auf Basiszeichen (z.B. a + ¨ = ä; aus zwei oder mehr Teilzeichen zusammengesetzte Zeichen in Hangul (Koreanisch))

Andere Zeichen: Grafikzeichen (Rahmen etc.), diverse Sonderzeichen (z.B. Währungssymbole)

Zeichenanzahl: 16-Bit-Kodierung; daher potentiell 65536 Codepunkte; im Moment umfaßt Unicode ca. 38000 Zeichen

Modelle für Lexikonstrukturen

(nach [Ide et al. 93]: Outline of a model for lexical databases. In: Information Processing & Management. 29(2).)

1. Textmodelle
   - Typographische Formatierung: die Marken enthalten Verarbeitungskommandos
   - Beschreibende Formatierung: die Marken enthalten Inhaltsinformation
   - Grammatische Formatierung: die Marken enthalten Inhaltsinformation; ihre Struktur wird durch eine Grammatik festgelegt
2. Relationale Modelle (keine Schachtelung der Attribute)
   Probleme:
   - Fragmentierung der Daten (Wörterbucheinträge sind sehr unterschiedlich)
   - die 'natürliche' hierarchische Struktur eines Eintrags ist nicht möglich
3. Relationale Modelle ohne Normalisierung (mit Schachtelung der Attribute)
   Probleme:
   - rekursive Schachtelung nicht möglich
   - Ausnahmen von der normalen Lexikonstruktur müssen auch hier gesondert behandelt werden.

Abgrenzung zwischen Lexikographie und Terminographie

nach Bess92, Terminologiekurs, Ecole de Traduction et d'Interprétation, Genève

Terminologie

bezeichnet die Gesamtheit der Fachwörter, die einer Kunst, einer Technik, einer Wissenschaft, einer Disziplin, einem Sachgebiet, einer Tätigkeit oder einer Praxis, einer Fabrik, einem Unternehmen, einer Schule, einem Wissenschaftler oder einer Gruppe von Wisenschaftlern, einem Autor eigen ist.

Bsp.: Terminologie der Informatik, der Landwirtschaftsmaschinen, der Chemie ...

Gegenstand

Lexikographie	**Terminographie**
Beschreibung der Gemeinsprache	Beschreibung von Fachsprachen
Tätigkeit im allgemeinen einsprachig	Tätigkeit im allgemeinen mehrsprachig
Erarbeitung von Wortsammlungen	Erarbeitung von Fachwortsammlungen

Ziele

Lexikographie	**Terminographie**
Zielpublikum: alle Sprechenden	Zielpublikum: Fachleute
Neigung zu einheitlichen "Produkten"	Vielzahl der Produkte (was Darstellung und Methoden angeht)
Vorwiegend Papierprodukte	Viele EDV-Produkte

Vorgehen

Lexikographie	**Terminographie**
Semasiologisches Vorgehen	Onomasiologisches Vorgehen
Feststellung des Gebrauchs	Normende Rolle
Diachronische und synchronische Beschreibung	Rein synchronische Beschreibung

Beschreibungsart

Lexikographie	**Terminographie**
Auszuwertender Korpus beliebig zusammengesetzt	Auszuwertender Korpus nach strengen Kriterien zusammengesetzt
Versuch einer Beschreibung der gesprochenen Sprache	Beschreibung der geschriebenen Sprache
Häufigkeit des Vorkommens für die Auswahl der Wörter wichtig	Suche nach seltenen Fachwörtern
Vorsicht gegenüber Neologismen	Besonderes Interesse für Neologismen
Vollständige Beschreibung des Wortschatzes unmöglich	Vollständige Beschreibung des Fachwortschatzes möglich

Art der Einträge

Lexikographie	**Terminographie**
Einfache Einträge	Mehrworteinträge
Vorhandensein aller grammatischen Kategorien	Überwiegend nominale Formen
Polysemie	Polyseme Fachwörter werden als Homonyme behandelt

Gegebene Information

Lexikographie	**Terminographie**
Phonetische Informationen	Keine Phonetik
Grammatische Informationen	Wenig grammatische Informationen
Etymologische Informationen	Keine etymologische Informationen
Historische Informationen	Keine historische Informationen

Beschreibungsmethode

Lexikographie	**Terminographie**
Sprachliche Definition	Definition durch Beschreibung des Gegenstands und des Begriffs
Beispiele aus der Literatur	Definitorischer und enzyklopädischer Kontext
Querverweise auf Synonyme, Antonyme	Informationsbringende Querverweise mit Assoziationen zu Gegenständen oder Begriffen, weniger zu Fachwörtern

Norm

Lexikographie	**Terminographie**
Gesellschaftliche und kulturelle Norm	Normung der Gegenstände, des Begriffs und der Fachwörter
Schwierige Orientierung des Gebrauchs	Freie Benennung
Gewicht des Sprachsystems	Gewicht der Normung

Autoren

Lexikographie	**Terminographie**
Linguisten	Fachleute (Übersetzer, Sachgebietsexperten, usw.); multidisziplinäre Arbeit

Gemeinsamkeiten

- Alphabetische Ordnung
- Beschriebene Elemente gehören dem Sprachsystem an (Wörter oder Fachwörter)
- Ähnlichkeit in der Methodologie

Gerold Schneider, Martin Volk