Korpuslinguistik
Anwendungen I: Lexikographie

Heike Zinsmeister
Sommersemester 2003
Universität Konstanz
FB Sprachwissenschaft
Raum G 116, Tel. 88 25 87, Fax. 88 48 65
Heike.Zinsmeister@uni-konstanz.de

24. Juni 2003

Textgrundlage

- Computergestützte Lexikographie
  Heid, Ulrich (2001). Computergestützte Lexikographie. In: Carstensen, Kai-Uwe; Ebert, Christian; Endriss, Cornelia; Jekat Susanne; Klabunde, Ralf & Langer, Hagen (Hrsg.) *Computerlinguistik und Sprachtechnologie. Eine Einführung.* Spektrum Akademischer Verlag, Heidelberg, 418-424.

Computergestützte Lexikographie

- Lexikon = Wörterbuch
  - Inventar von Wörtern, Wortbestandteilen, Wortgruppen
  - (verarbeitungs)relevante Informationen zu den Wörtern etc.
- lexikographische Arbeit
  - (zeit-, kosten-) aufwändig
  - relativ großer Anteil Routinearbeit

  → Arbeitserleichterung durch Computer
- Computerunterstützung
  - lexikalische Akquisition
  - Repräsentation linguistischer Daten

Lexikographie

- Erstellung von Wörterbüchern (WBn)
- wissenschaftliche Beschäftigung mit WBn
  - WB-Analyse
- Metalexikographie (Hausmann 1985)
  - WB-Kritik
  - Theorie der Lexikographie
  - Geschichte der Lexikographie
  - Forschung zur Benutzung

Grundbegriffe zum Wörterbuchaufbau

- der Hauptteil eines WB besteht aus Artikeln
- Artikel = Lemma/Stichwort + dazugehöriger Information
- Makrostruktur = geordnete Menge aller Lemmata
- Mikrostruktur = gesamte Informationen zu einem Lemma in einem Artikel
  - erklärende Information: Definition
  - syntagmatische Information: Konstruktionen, Kollokationen
  - paradigmatische Information: Synonyme, Antonyme, Homonyme, Paronyme, Wortbildung

(zusammengefasst aus: http://www.uni-stuttgart.de/lingrom/stein/kurs/lexikogra

Wörterbucherstellung

- Format
  - einsprachig
  - zweisprachig
  - mehrsprachig
- Medium
  - gedruckt
  - (interaktiv) am Rechner nutzbar
  - maschinenlesbar für NLP-Anwendungen

Lexikographische Arbeit

- linguistische Datenbeschreibung
  - Lexikon-Spezifikation: Informationstypen, Klassifikationskriterien
  - Einträge: Feststellen, Auswählen, Klassifizieren von relevanten Wort-Eigenschaften
- didaktische Präsentation der Artikelinhalte (explizite versus implizite Darstellung)
  - visuell (Symbole, Text)
  - typographisch (kursiv, fett, Klammerung, Einrückungen)
  - textuell (Beispielsätze, Zitate)

Akquisition lexikalischer Information

- Beschaffung von Daten
- Quellen für traditionelle Arbeit
  - Textmaterial
  - Belegkarteien
  - elektronische Versionen früherer Ausgaben eines WB
  - Sammlung mehrerer WB
  - Intuition des Redakteurs
- (zusätzliche) Quelle bei Computerunterstützung
  - Korpora

Akquisition aus Textkorpora (1)

- Identifikation von Beispielsätzen, Klassifikation
- Wissen über Wörter ableiten
- Verfügbarmachung von Satz, Phänomen und Wort(gruppe)
- Hilfsmittel: Konkordanz (KWIC)
  - Kontextangabe zu Vorkommnissen von Wort, Lemma, regulärem Ausdruck (Suchmuster)
  - Lexikograph leistet linguistische Beschreibung und Interpretation

Akquisition aus Textkorpora (2)

```
# Context: 25 characters left, 25 characters right
#
# Query: BNC-SAMP; Q1=[word="[Ll]ov.*"][pos="AT.*"][pos="N.*"];
#------------------------------------------------------------------------------
   329949: tamine tablets . I would  <love a perm> as my daughter is gettin
   471506: are too complex . If you  <love the sea> tackle your concerns hea
  1128348: be much of a poem But I   <love the title> . <pb_n=1> I <pb_n=2>
  1625861: al love interest . Essex  <loves the Queen> since she is the embodim
  1893783: y eyesight . No . No . I  <love the Rocky> Horror <ptr_t=KBULC0FD>
  1922676: . Oh <ptr_t=KC0LC0R4> I   <love the jacket> ! <ptr_t=KC0LC0R5>

  1983255: know there , <unclear> I  <love the way> that pass both passports
  2363140: No way . <pause> I just   <love the way> <pause> that kid must 'v
  2402975: licks_tongue"> <pause> I  <love the way> she does that , she goes
```


Akquisition aus Textkorpora (3)

- Nutzung von erschloßenen, annotierten Korpora
- Korpuserschließung
  - Tokenisierung
  - Wortklassen-Tagging
  - Lemmatisierung
  - Chunking
  - Parsing

Akquisition aus Textkorpora (4)

- Schema
  1. Aufbereitung und linguistische Annotation der Korpora
  2. Extraktion von Wörtern, Wortkombinationen, Phrasen, Sätzen mittels musterbasierter Suche (Einbezug der Annotation), ggf. spezielle Suchwerkzeuge; ggf. statistische Analyse (Bestimmung von Verteilung, relative Häufigkeit, Signifikanz)
  3. Interpretation der Ergebnisse bzgl. Klassifikation des Ziel-WBs; (interkativer) Einbau ins Ziel-WB

Akquisitionsziele (1)

- syntaktische Subkategorisierung
  (sehr wichtig für NLP-Anwendungen)

- Distribution
  - aktiv versus passiv bei Verben
  - prädikativ vs. adverbial vs. attributiv bei Adjektiven
  - adnominal vs. adverbial vs. adadjektivisch bei Adverbien
  - definit vs. indefinit bei Nomina

- Kollokationen
  - *Pause einlegen, helle Aufregung*

Akquisitionsziele (2)

- lexikalisch-semantische Relationen (Taxonomie)
  - Antonym, Hyponym, Hyperonym, Meronym, etc.

- lexikalisch-semantische Klassen
  - Verbale Klassen, wie *Bewegungsverben, Zustandsänderungsverben, Verba Dicendi*
  - Nominalisierungen, wie *Ereignis- versus Objektslesart*

- Ausblick
  - Äquivalenzkandidaten aus parallelen Korpora (mehrsprachig, satz- oder abschnittsweise zugeordnet)

Akquisition aus traditionellen WBn (1)

- Analyse von elektronischen Fassungen gedruckter WB
  (erstmals 1980)

- Extraktion linguistischer Information für NLP-Anwendungen
  - morphologische
  - syntaktische
  - Definitionen (“*ist ein . . . , besteht aus, . . .* ”)
    $\rightarrow$ Begriffshierarchien, Ontologien
    Regel: ’genus proximum differentia specifica’
  - Methoden: syntaktische Analyse der Definitionstexte, musterbasierte Suche

Akquisition aus traditionellen WBn (2)

- Nutzung für WB-Erstellung
  - Analyse des Quell-WBs:
    deskriptives Programm und Präsentationsmittel
  - Definition der Abbildung:
    Beschreibung Quell-WB $\rightarrow$ Klassifikation Ziel-WB
  - Extraktion der Angaben
    aus Quell-WB, Reformatierung, Einbindung in Ziel-WB

Repräsentation (1)

- inhaltliche Aspekte
  - Auswahl linguistischer Information
    - Klassifikation
    - Kodierung
    - Zugreifbarkeit
  - angestrebte Anwendungen des Lexikons
    - PoS-Tagging
    - Indexierung im Information Retrieval
    - Maschinelle Übersetzung
  - linguistische Theorie, Beschreibungsansatz
    (ist multifunktionales Lexikon möglich? Vgl. IDS Mannheim)

Repräsentation (2)

- technische Aspekte
  - Wahl des Räpresentationsformat/-formalismus
    - Datenbank
    - Textdatei mit Markups (z.B. XML, SGML)
    - Datenstruktur computerlinguistischer Formalismen
      (z.B. Merkmalsstruktur, typisierte Merkmalslogik)
  - Abfrage
  - Reformatierbarkeit für Anwendungen
  - Anbindungen an Interfaces für Datenpflege
- Standardisierung (z.B. EAGLES 2001)
  - Inhalt und Form der WB
  - Informationstypen für NLP-Lexika: Morphologie, Morphosyntax, Subkategorisierung

Lexikographiesysteme

- Lexikographen-Arbeitsplatz (vgl. Atkins 1992, Heid et al 2000)
  - unterstützt Erstellung und Überarbeitung von WBn
  - spezifisch Editorfunktion (idealerweise DTD-basiert)
  - Zugriff auf die Akquisitionsergebnisse
    - Korpusanalyse
    - korpusstatistische Angaben
    - Einträge oder Beispiele aus früheren Versionen des WBs oder anderen WBn desselben Verlags
    - Daten aus Zitatsammlungen
    - direkte Korpusrecherche mit KWIC-Werkzeugen

Lexikographenarbeitsplatz

- interaktive Arbeit mit lexikalischer Information
  - Akquisition
  - Repräsentation
  - Anwendung
  - Korpuszusammenstellung

Automatische Exzerption

- Überarbeitung bestehender WB (Makro- und Mikrostruktur: Eintragswörter, Kollokationen, Grammatikregeln, Beispiele, etc.)
  - Vergleich von verschiedenen WBn
  - Vergleich von WB mit Korpusdaten
  - Ergänzungsvorschläge (’Aufnahmekandidaten’)
  - Löschvorschläge (’Löschkandidaten’)
  - bestehende Systeme
    - Transferbereich 32 (TFB 32), IMS Stuttgart
    - A Semi-Automatic Lexicographer’s Workbench for Writing **W**ord **S**ense **P**rofile**S** (WASPS), ITRI Brighton

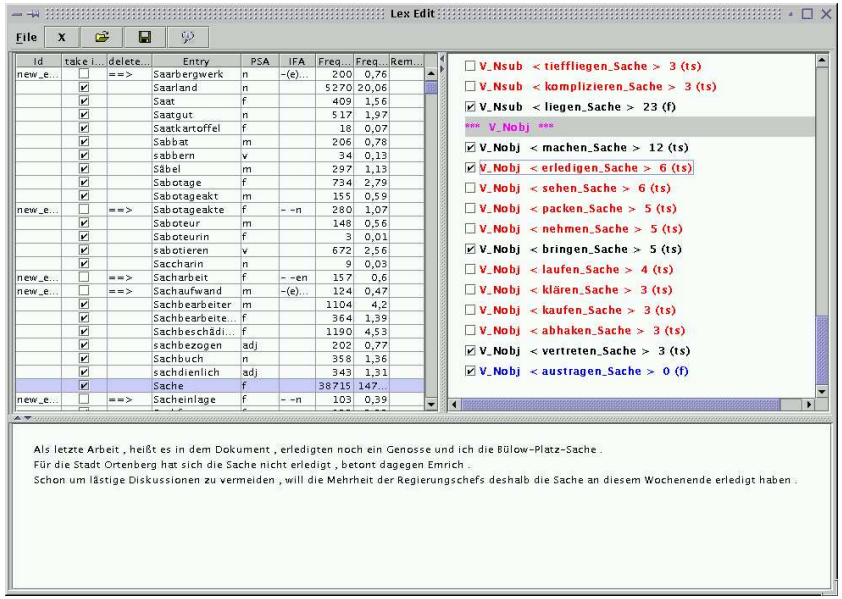

Figure 2: Bildschirmpräsentation der Suche nach *Sache*, TFB 32

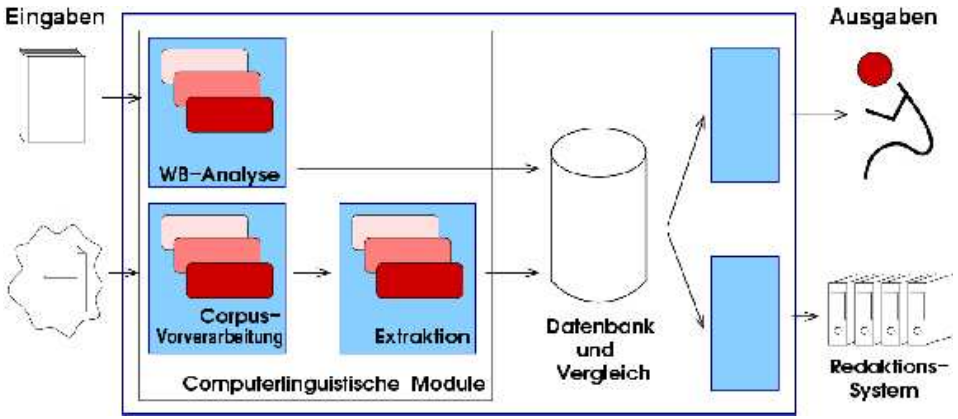

Figure 1: Architektur: Lexikographenarbeitsplatz TFB 32

Computational Terminology

- Terminologie: Fachwortschatz-Arbeit
- computergestützte Akquisition, Bearbeitung, Repräsentation, Standardisierung von Fachwortschatz
  - Extraktion von Termkandidaten aus Texten, nach linguistischen Kriterien, mit statistischen Verfahren
  - Akquisition von Übersetzungvorschlägen aus parallelen Texten (z.B. Wetternachrichten, Automobilindustrie)
  - Akquisition von definitionsrelevantem Material, Belegen für ontologische Relationen
  - Repräsentation multilingualer terminologischer Daten, Aufbau von Thesauri und Wissensbanken
  - Terminologische Standardisierung innerhalb von Unternehmen, Definition von kontrolliertem Vokabular

Zusammenfassung und Ausblick

- Einsatz von computerunterstützter Lexikographie
  - Universitäten
  - Verlage
  - Dokumentationsfirmen
  - Übersetzungsfirmen
  - jegliche Softwarefirma, die mit computerlinguistischen Methoden arbeitet braucht elektronische WB
- Verfügbarkeit von annotierten Korpora, Baumbanken wird bessere Unterstüzung ermöglichen
- Experimente zum Einsatz von Verfahren des maschinellen Lernens für die Lexikonakquisition

Referenzen

- Atkins, Sue (1992) Tools for computer-aided corpus lexicography: the Hector project", in Papers in Computational Lexicography: Complex'92, F. Kiefer, G. Kiss and J. Pajsz (eds.) Hungarian Academy of Sciences, Budapest. pp. 1-60. Also in Acta Linguistica Hungarica 41, F. Kiefer (ed.) (1991), Hungarian Academy of Sciences, Budapest: Akadémiai Kiadó.
- Hausmann, Franz Josef (1985): Lexikographie. In: Schwarze, Christoph/Wunderlich, Dieter (Hgg.): Handbuch der Lexikologie. Berlin, 368-377.
- Heid, Ulrich; Evert, Stefan; Docherty, Vincent; Worsch, Wolfgang and Wermke, Matthias (2000) A data collection for semi-automatic corpus-based updating of dictionaries in Ulrich Heid; Stefan Evert; Egbert Lehmann and Christian Rohrer, editors, Proceedings of the 9th EURALEX International Congress, 183 – 195.(http://www.ims.uni-stuttgart.de/ uli/papers/elx99-abs.ps.gz)
- Kilgarriff, Adam und David Tugwell (2001) "WORD SKETCH: Extraction and Display of Signifi cant Collocations for Lexicography". In Proc. workshop "COLLOCATION: Computational Extraction, Analysis and Exploitation", 32-38. 39th ACL & 10th EACL, Toulouse (http://www.itri.bton.ac.uk/∼David.Tugwell/colloc.ps)
- Transferbereich 32: http://www.ims.uni-stuttgart.de/projekte/TFB/TFB-Webseite
- WASPS: http://wasps.itri.bton.ac.uk/

Fragenkatalog

- Was ist musterbasierte Suche? Erläutern Sie das Vorgehen anhand von Beispielen und nennen Sie ein Anwendungsgebiet.
- Beschreiben Sie einen Lexikographen-Arbeitsplatz. Welchem Zwecke dient er?
- Was versteht man unter 'Automatischer Exzerption'?
- Nennen Sie mögliche Akquistionsziele.
- Was ist 'Computational Terminology'?