Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Arbeitsstelle „Computerphilologie“

Sprachressourcen in der Lehre: Erfahrungen aus der historischen Korpuslinguistik

Cristina Vertan
cristina.vertan@uni-hamburg.de

Überblick

- Das Kontext (Lehrangebot an der Universität Hamburg)
- Das Szenario (Ziele des Seminars)
- Benutzte Corpora
- Ergebnisse (Erfolge, Hindernisse)
- Zukunftsaktivitäten

Das Kontext

- Vorlesungen und Seminare mit Bezug auf älterer Deutsche Sprache und Literatur sind Teil des :
  - BA-Studiengangs „Deutche Sprache und Literatur“ und
  - MA-Studiengangs „Germanistische Linguistik und allgemeine Sprachwissenschaft“
- Kein Master- oder Bachelor-Studiengang im Computer- oder Korpuslinguistik
- Unregelmässig gibt es Veranstaltungen mit Bezug auf Computerlinguistik oder Computerphilologie, meistens ohne praktischer Anteil

Beispiele von Seminare unv Vorlesungen mit Computerlinguistischen Bezug

- Vorlesung „Computerphilologie“ - ohne praktische Übungen
- Seminar II „Linguistische Probleme der Maschineller Übersetzung“
- Seminar II „Wort, Name, Begriff, Terminus. Ihre linguistische Theorie und ihre Relevanz für computergestützte Anwendungen“
- Seminar Ib - „Kohärenz und ihre computerlinguistische Beschreibung

Hintergrundkenntnisse von Studenten

- Grundkenntnisse in der Linguistik, allerdings nicht homogen, da die „Einführung in der Linguistik“ von unterschiedlichen Dozentenangeboten wird.
- Einige Studenten haben andere Veranstaltungen über ältere Deutsche Sprache und Literatur besucht.
- Meistens Studenten haben elementaren Kenntnisse des Umgangs mit Rechner (Internet-Recherche, MS-Office)
- Einige wenige Studenten haben theoretische Kenntnisse über computergestützte Morphologie oder Syntax

Überblick

- Das Kontext (Lehrangebot an der Universität Hamburg)
- Das Szenario (Ziele des Seminars)
- Benutzte Corpora
- Ergebnisse (Erfolge, Hindernisse)
- Zukunftsaktivitäten

Das Szenario

- Seminar II „Korpuslinguistik“ -2 SWS - (Prof. Dr. Renata Sczepaniak) + Übung „Übung zu Korpuslinguistik“ -2 SWS- (Dr. Cristina Vertan)
- Starke Orientierung auf diachronen Aspekten
- Die aktive Teilnahme an der Übung ist Bestandteil des Scheins für das Seminar.
- Die Studenten schreiben eine gemeinsame Hausarbeit (für Seminar und Übung) in dem Sie obligatorisch für die linguistische Untersuchung mindestens eine computergestützte Methode anwenden müssen.

Berabeitete Themen aus der Computerlinguistik

- Kodierung; Zeichensätze
- Reguläre Ausdrücke
- Elementare Programmierung in Perl
- Statistische Grundlagen für Korpuslinguistik
- Kollokationen, PoS Tagging, Syntaktische Dependenzen
- Der Auswahl dieser Themen wurde in der ersten Sitzung begründet.

Begründungsbeispiel -1-
Was ist zu beachten bei einem Korpus in elektronischer Form?

- Meistens existierenden Korpora in elektronischer Form sind durch:
  - Digitalisierung = OCR + manuelle Korrektur oder
  - Abschreiben (Eintippen)

  von Dokumenten im Papierform
- Wichtig für eine sinnvolle weitere Benutzung des Korpus ist die Auseinandersetzung mit folgenden Problemen:
  - Im welchen Format muss das Korpus gespeichert wird (Bild, Text) ?
  - Was für Zeichensätze benötigt werden
  - Wie wird das Korpus visualisiert?

Thema : Kodierung, Zeichensätze

Begründungsbeispiel -2-
Wie kann ich in einem Korpus suchen?

- Eintippen in den Suchfeldern von einzeln Wörtern
  - Ergebnis: Im Korpus werden alle Plätze wo die Wörter erscheinen (zusammen oder getrennt) gefunden
- Benutzung von s.o.g. „erweiterte Suche“, wo logische Operatoren (AND, ODER, NOT) und/oder spezielle Charakteren (z.B. *,+) verwendet werden können
- Die Möglichkeiten bei der „erweiterter Suche“ unterscheiden sich von Suchmaschine zur Suchmaschine, d.h:
  - **Man muss im voraus, welche Operatoren und spezielle Charakteren zugelassen sind und was für Bedeutung diese spezielle Charakteren haben, überprüfen**
- Die Kombination von Suchbegriffen, spezielle Charakteren und logische Operatoren nennt man: **regulären Ausdruck**

Thema : reguläre Ausdrücke

Begründungsbeispiel -3-
Wie kann ich Korpuseigenschaften berechnen ?

Die **Korpuslinguistik ist ein Bereich der Linguistik, in dem Theorien über Sprache anhand von Belegen oder statistischen Daten aus Textkorpora aufgestellt oder überprüft werden. (*Wikipedia*)**

- Die Zählung vom Auftreten bestimmten Phänomenen (z.B. Anzahl von unterschiedlichen Wörtern, Anzahl von Eigennamen, Anzahl von Sätzen) sowie
- Die Erstellung und Speicherung von Listen die diese Phänomenen beinhalten

Kann automatisch mit Hilfe von einfachen Programmierungsbefehlen erfolgen

Thema : einfache Programmierung im PERL

Begründungsbeispiel -4-
Was kann man in einem Korpus berechnen?

- Bekannteste Berechnung ist die Frequenz mit der einen Wort, einen Mehrwortausdruck oder einen linguistischen Phänomen vorkommt.
- Neben Frequenz gibt es auch andere Messungen die aussagekräftiger über den Korpus-Struktur sind wie z.B.: Durchschnitt und geometrischen Mittel.
- Manchmal neben Frequenz muss man auch die Signifikanz messen (ob ein Wort 50 Mal in einem 1000-Token Korpus ist anders wenn derselben Wort in einem 100 000 –Token Korpus erscheint

Thema : statistische Berechnungen in einem Korpus

Begründungsbeispiel -5-
Wie kann man (linguistische) Merkmale im Korpus speichern ?

- Um die beobachtete Phänomenen langfristig zu speichern und die Daten austauschbar zu machen braucht man einen Standard Format.
- Als Standard Format hat sich XML (eXtensible Mark-up Language) etabliert.
- Die Markierung erfolgt durch <Tags> die man definiert oder aus einem vordefinierten Bibliothek auswählt.
- Die Tags folgen einer, mit Hilfe von XML-Schema, fest gelegter Syntax.
- Öfter sind Korpora bereits annotiert und die Tags werden selbst für Suche oder statistische Berechnungen benutzt

Thema : Annotierung im XML

Begründungsbeispiel -6-
Höhere Korpusverarbeitung

- Oft untersucht man in einem Korpus nicht nur den Auftritt einzelnen Wörter sondern auch
  - Der gesamten Auftritt von morphologischen Varianten des Wörtes (z.B. Flexionsformen)
  - Kohärenz-Phänomenen
  - Syntaktischen Umfeld
  - Kollokationen

Thema : höhere linguistische Bearbeitung

- Dafür braucht man Werkzeuge wie :Lemmatiser, PoS Tagger, Parser
- Diese automatische Werkzeuge haben immer eine Fehlerquotedie einerseits bekannt und anderseits verstanden werden muss
- Für historische Sprachen sind diese Werkzeuge mit viel Vorsicht anzuwenden

Aufbau des Seminars -2-

- Jedes Thema wurde in 2-3 Sitzungen bearbeitet
  - In der erster Sitzung wurden die Problemen und die möglichen Lösungen dargestellt
  - In der nächster Sitzung wurden mit Hilfe dieser Lösungen eine Fragestellung aus dem Seminar „Korpuslinguistik“ abarbeitet
- (Für jedes Thema wurden kleine Hausarbeiten verteilt)

Ablauf des Seminars -3-

1 Themenüberblick; Korpus-Kodierung und Speicher-Formate
2 Reguläre Ausdrücke
3 Anwendung von Reguläre Ausdrücke für die Untersuchung von der Verteilung von Genitiv-Bildung
4 Elementare Programmierung im PERL
5 Anwendung von Programmierungssprache Perl bei der Untersuchung von Substantivgroßschreibung
6 Statistische Grundlagen für Korpuslinguistik
7 Anwendung von Statistische Messungen für „Genitiv-Bildung“ und „Substantivgroßschreibung“
8 XML-Grundlagen
9 Kollokationen, PoS Tagging, Syntaktische Dependenzen
10 Untersuchungen für die Verwendung von dher/ther/der
11 Verteilung und Anwendung des „-in“ Suffixes
12 Verteilung und Anwendung des „-in“ Suffixes

Überblick

- Das Kontext (Lehrangebot an der Universität Hamburg)
- Das Szenario (Ziele des Seminars)
- Benutzte Corpora
- Ergebnisse (Erfolge, Hindernisse)
- Zukunftsaktivitäten

Benutzte Korpora

- Cosmas Korpus(http://www.ids-mannheim.de/cosmas2/)
  - für das Thema „Genitiv Bildung“
- Das Bonner Frühneuhochdeutsch-Korpus (http://korpora.zim.uni-duisburg-essen.de/Fnhd/)
  - Für das Thema “Substantivgroßschreibung“
- TITUS Korpus (http://titus.uni-frankfurt.de/indexd.htm)
  - benutzte Texte: Tatian, Otfried, Notker
  - Für das Thema „Definitartikel in althochdeutsch“
- DWDS Kernkorpus (http://www.dwds.de/)
  - Für das Thema „Untersuchungen an der Benutzung des –in Suffixes“

Aufgabe-Beispiel

- Laden Sie aus AGORA das Test -Korpus DWDSTest
- Schreiben Sie ein Programm, der in diesem Korpus alle Wörter, die mit **Ab-, Rück-, Wieder-, Zu-, Um-, Vor-, Ur-,, Vol**l-anfangen und mit **-es** enden und schreiben Sie die Ergebnisse in einer Datei.
- Verbessern Sie das Programm so, dass unvernünftige Ergebnisse vom Anfang an nicht mehr gespeichert werden (Erstellen Sie einen Array mit Stop-Wörtern).
- Erstellen Sie die Types und deren Anzahl
- Wiederholen Sie das Ergebnis mit denselben Präfixen und Endung -**s** und ohne **e**.
- Vergleichen Sie jetzt die beide Ergebnisse: erzeugen Sie eine einzige Datei und sortieren Sie mit Hilfe von Perl die Einträge

Überblick

- Das Kontext (Lehrangebot an der Universität Hamburg)
- Das Szenario (Ziele des Seminars)
- Benutzte Corpora
- Ergebnisse (Erfolge, Hindernisse)
- Zukunftsaktivitäten

Erfolge

- Die Studierende könnten am Ende des Seminars mindestens einfache regulären Ausdrücke in der Suchfeldern der o.g. Korpora zu formulieren
- Sie haben gelernt, wie man mit unterschiedlichen Zeichen umgeht und waren in der Lage kleine Hilfsprogrammen in Perl zu erfassen.
- Sie konnten bewerten die Nützlichkeit von computergestützte Methoden in der Korpuslinguistik.

Hindernisse

- Jedes Korpus benutzt seiner eigener Variante von Kodierung von Regulären Ausdrucken. Daher ist es schwierig besonders für ungeübten Benutzern einen Überblick zu behalten.
- die Web-Version des COSMAS-Korpus war bei gleichzeitige Benutzung von etwa 12 Studenten ziemlich langsam
- Für die Programmierübungen ist es extrem wichtig das die Studierenden Teile des Korpus selbst manipulieren können. Das ist meistens momentan nicht möglich
- Für 24 Stunden /Semester war das CL-theoretischen Hintergrund doch zu anspruchsvoll. Die Programmierungsanteil benötigt viel mehr Übungen

Überblick

- Das Kontext (Lehrangebot an der Universität Hamburg)
- Das Szenario (Ziele des Seminars)
- Benutzte Corpora
- Ergebnisse (Erfolge, Hindernisse)
- Zukunftsaktivitäten

Zukunftsaktivitäten

- Das System: Linguistische Seminar + computerlinguistische Übung, basiert auf einem linguistischen Thema soll fortgesetzt werden
- Allerdings müssen die computerlinguistische Verfahren eingeschränkt werden z.B.
  - reguläre Ausdrücke +Perl oder
  - statistische Verfahren + Perl
  - oder XML-Annotierung + Auswertung
- Es wäre sehr hilfreich wenn für Lehrzwecken, Teile der o.g. Corpora zumindest im Text-Form (Unicode) verfügbar gemacht werden können