Exposé - Studienarbeit Falko R*

Karsten Hütter
huetter@informatik.hu-berlin.de

Betreuer:

Prof. Ulf Leser
Lehrstuhl für Wissensmanagement in der Bioinformatik†

Prof. Anke Lüdeling
Institut für deutsche Sprache und Linguistik, Korpuslinguistik‡

23. April 2007

*Falko R: Falko mit relationaler Datenspeicherung
†HUB WIB: http://www.informatik.hu-berlin.de/forschung/gebiete/wbi/
‡HUB Korpuslinguistik: http://www2.hu-berlin.de/korpling/

1 Hintergrund

FALKO ist ein fehlerannotiertes Lernerkorpus des Deutschen. Es enthält Texte von fortgeschrittenen Lernern des Deutschen als Fremdsprache. Die EXMARaLDA-basierte Datenhaltung ermöglicht es, Korpusdaten mit konfligierenden Annotationsebenen abzubilden [SLM06].

Die Studienarbeit soll überprüfen, ob das bestehende System durch ein relationales Datenbanksystem und ein geeignete Programmierschnittstelle ersetzt werden kann.

Die Betrachtungen schließen eingeschobene Annotationen(Abschnitt 1.2), Indexierungsmöglichkeiten relationaler Datenbanksysteme und eine modulare Software Schnittstelle ein.

In Abgrenzung dazu stehen Performanzaspekte für die geringen Datenmengen in FALKO im Hintergrund. Diese könnten in weiterführenden Arbeiten thematisiert werden.

1.1 Datenhaltung

FALKO verwendet das multi-layer EXMARaLDA Datenformat, ein XML Format, dass zur Speicherung von Gesprächsdaten entwickelt wurde [SW05]. Die Daten werden in die IMS Corpus Workbench CWB überführt und durchsuchbar gemacht [CSHK99]. Da die CWB für flach annotierte Tabellenkorpora entwickelt wurde, gehen bei der Konvertierung strukturelle Informationen der Annotationen verloren.

Die Universität Potsdam entwickelt im Projekt ANNIS eine Plattform zur Speicherung und Durchsuchung von Korpusdaten. ANNIS verfolgt einen hauptspeicherorientierten Ansatz [GD06].

Das British National Corpus BNC Projekt konzentriert sich auf die Speicherung großer, flach annotierter Daten geschriebener Sprache. Die Benutzerschnittstelle der BNC ist Xaira [bnc].

Desweiteren arbeitet die Universität Hamburg mit dem EXMaRALDA Suchwerkzeug Zecke über Gesprächsdaten [Sch].

1.2 Eingeschobene Annotationen

Die kleinste Dateneinheit in FALKO ist der Token. Token beschreiben Worte und Interpunktionen von Texten. Grundlegend wird hierbei ein Token über das geschriebene -oder transkribierte gesprochene- Wort charakterisiert. Zusätzlich werden die Wortart und das Lemma für jedes Token vollautomatisch annotiert. Wortarten und deren Festlegung werden in sog. Tagsets definiert. Falko verwendet das Stuttgart-Tübingen-Tagset STTS [STT95].

Moderne Korpusarchitekturen unterstützen Annotation mit unterschiedlichen Granularitäten. Sie bieten die Möglichkeit einzelne Token, Sätze, Texte oder beliebige Tokensequenzen zu annotieren. In der Fehlerannotation von Lernertexten gibt es den Spezialfall der eingeschobenen Annotation. Sie dient u.a. der Beschreibung von Wortauslassungen. Hierbei werden in der EXMaRALDA Abbildung des Textes leere Elemente auf der Hauptspur eingefügt und auf den Annotationsspuren beschrieben.

Abbildung 1 zeigt die eingeschobene Annotation A3.1. Das EXMARaLDA Format und seine Werkzeuge unterstützen diese Form der Annotationen durch Einfügen von neuen Items auf der Hauptspur. Dennoch würde der Import in die CWB durch die entstehenden Leertoken auf der Hauptspur zu nicht definierten Resultaten führen.

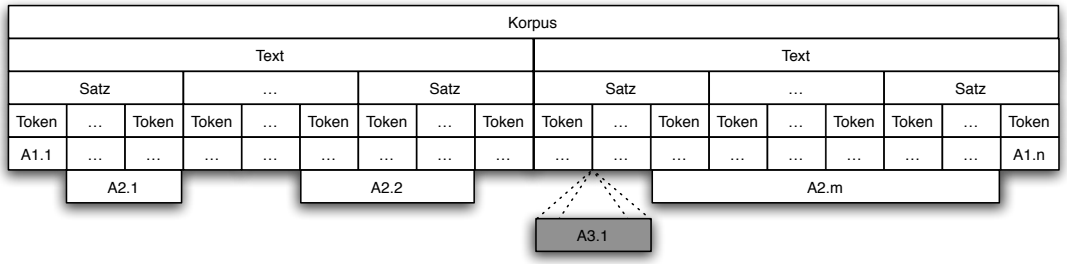

Abbildung 1: Schema eines mehrebenenannotierten Korpus mit ausgezeichneter eingeschobener Annotation A3.1.

2 Vorgehen

Erste Versuche mit der Transformation der Daten aus FALKO in das durch Abbildung 2 illustrierte Datenbankschema haben bereits gute Ergebnisse gezeigt. Diese müssen nun genaueren Betrachtungen standhalten.

Nach erfolgtem Datenimport werden die Abfragemöglichkeiten und Transformation einer Objektrepräsentation von Suchanfragen nach SQL untersucht. Ziel ist es dabei bestehende CWB-Funktionalität vollständig abzudecken. Es werden Vor- und Nachteile dieser Implementierung im Vergleich mit dem Projekt ANNIS diskutiert.

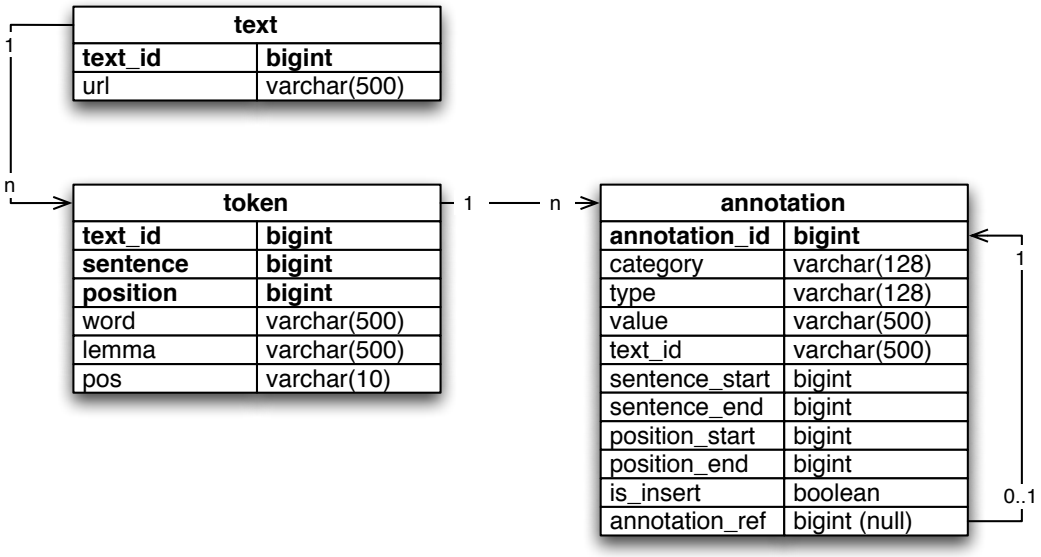

Abbildung 2: Naives relationales Datenschema für die Speicherung mehrebenenannotierter Korpora.

Literatur

[bnc] *About the British National Corpus*. `http://www.natcorp.ox.ac.uk/corpus/index.xml?style`

[CSHK99] Christ, Oliver ; Schulze, Bruno M. ; Hofmann, Anja ; König, Esther: *The IMS Corpus Workbench: Corpus Query Processor (CQP) - User's Manual*. `citeseer.ist.psu.edu/christ99ims.html`. Version: 1999

[GD06] Götze, Michael ; Dipper, Stefanie: *ANNIS: Complex Multilevel Annotations in a Linguistic Database*. `http://acl.ldc.upenn.edu/W/W06/W06-2709.pdf`. Version: 2006

[Sch] Schmidt, Thomas: *ZECKE - Ein Prototyp für ein Suchwerkzeug für EXMARaLDA-Daten*

[SLM06] Siemen, Peter ; Lüdeling, Anke ; Müller, Frank H.: *FALKO - Ein fehlerannotiertes Lernerkorpus des Deutschen*. 2006

[STT95] Schiller, A. ; Teufel, S. ; Thielen, C.: *Guidelines für das Tagging deutscher Textcorpora mit STTS*. `http://www.sfs.uni-tuebingen.de/Elwis/stts/stts-guide.ps.gz`. Version: 09 1995

[SW05] Schmidt, Thomas ; Wörner, Kai: Erstellen und Analysieren von Gesprächskorpora mit EXMARaLDA. In: *Ge-*

sprächsforschung - Online-Zeitschrift zur verbalen Interaktion 6 (2005). http://www.gespraechsforschung-ozs.de/heft2005/px-woerner.pdf