Anleitung zum Korpus DEREKO

1. Korpusaufbau/Metadaten

Das Deutsche Referenzkorpus (kurz DeReKo) ist ein elektronisches Archiv von deutschsprachigen Textkorpora, das seit 1964 existiert und vom Institut für Deutsche Sprache (IDS) in Mannheim gepflegt und kontinuierlich ausgebaut wird. Mit derzeit über 5 Milliarden laufenden Textwörtern (Stand: Februar 2012) ist DeReKo die weltweit größte Sammlung elektronischer Korpora der deutschen Gegenwartssprache, die für wissenschaftliche Zwecke bestimmt ist.

Das gesamte in COSMAS II (Suchprogramm für Korpora) recherchierbare DeReKo-Archiv setzt sich aus 108 Teilkorpora zusammen. Diese lassen sich wie folgt untergliedern: Zeitungs- und Magazintexte, Historische, v.a. literarische Texte, Gebrauchstexte, bestehende Korpora, morphosyntaktisch annotierte Texte usw.

Die 108 Teilkorpora sind in 10 Archiven organisiert: Archiv der geschriebenen Korpora, Archiv morphosyntaktisch annotierter Korpora (CONNEXOR-*Tagset*), Archiv morphosyntaktisch annotierter Korpora (TreeTagger-*Tagset*), Archiv morphosyntaktisch annotierter Korpora (MECOLB-*Tagset*), Archiv der historischen Korpora, Archiv des Projekts "Sprachliche Umbrüche des 20. Jahrhunderts", Korpus-Kartei der Gesellschaft für deutsche Sprache, Archiv der phasengegliederten Wendekorpora, Archiv der aussortierten geschriebenen Korpora, und das Archiv der *Süddeutschen Zeitung*. Archivübergreifende Suchanfragen sind nicht möglich.

Das Deutsche Referenzkorpus ist nicht als ein ausgewogenes Korpus, sondern als eine Art Ur-Stichprobe des Sprachgebrauchs konzipiert, aus der sich jeder Nutzer selbst ein virtuelles Korpus zusammenstellen kann, das für seine spezielle Fragestellung angemessen ist und für die von ihm untersuchte Sprachdomäne repräsentativ ist.

Im DEREKO sind folgende Metadaten enthalten: Veröffentlichungsdatum, Entstehungszeit, Autorenname, Herausgeber, Publikationsmedium, Publikationsort, Textsorte, Themenkategorie, Duplizitätsinformationen, quantitative Angaben zu Wörtern, Sätzen, Absätze, Anzeichen für neue/alte Orthografie, Lizenzbedingungen.

2. Morphosyntaktische Annotationen

Ein Teil der Texte aus Archiv W liegt mittlerweile mit morphosyntaktischen Annotationen aufbereitet vor. Die Annotationen wurden von verschiedenen Taggern generiert und liegen unterschiedlichen Tagsets zugrunde. *Tagsets* und *Tagger* wurden unabhängig voneinander außerhalb des IDS entwickelt und lassen sich wenig bis gar nicht aufeinander abstimmen. Wegen des damit verbundenen enormen manuellen Aufwandes werden Annotationen nicht überprüft bzw. korrigiert.

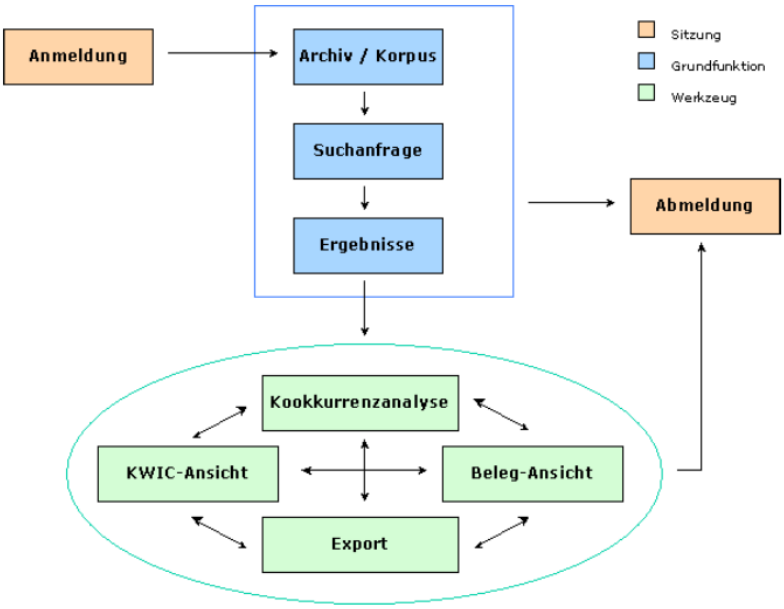
Archiv	**Tagset**	**Tagger**	**aktualisiert**
TAGGED-C		CONNEXOR	2010
TAGGED-T	STTS	TreeTagger	2011
TAGGED-M	MECOLB	LOGOS	1998

Der Treetagger ist ein Werkzeug für automatische Annotationen von Texten mit part-of-speech und Lemma Informationen.

3. Kurze Anleitung zur Benutzung

Eine Recherche in COSMAS II ist wie folgt aufgebaut (Struktogramm auf

http://www.ids-mannheim.de/cosmas2/projekt/hilfe/quick.html):



Online-Hilfe zu COSMAS II[web]:

http://www.ids-mannheim.de/cosmas2/web-app/hilfe/allgemein/

4. kommentierte Literaturhinweise

Bopp, Sebastian (2010): Einführung in die Korpuslinguistik mit DeReKo und COSMAS II. <www.philhist.uni-augsburg.de> [abgerufen am 03.11.2012]

Bubenhofer, Noah (2006): Einführung ind die Korpuslinguistik: Praktische Grundlagen und Werkzeuge. <http://www.bubenhofer.com/korpuslinguistik/kurs/index.php?id=cosmas_client_annotiert.html> [abgerufen am 03.11.2012]

Kupietz Marc/Belica Cyril/Keibel, Holger/Witt, Andreas (): The German Reference Corpus DeReKo: A Primordial Sample for Linguistic Research http://www.lrec-conf.org/proceedings/lrec2010/pdf/414_Paper.pdf [abgerufen am 03.11.2012]

IDS: Deutsches Referenzkorpus/Archiv der Korpora geschriebener Gegenwartssprache <http://www.ids-mannheim.de/cosmas2/uebersicht.html> [abgerufen am 05.11.2012]