Institut für Informatik der Universität Zürich
Abteilung Computerlinguistik
Prof. Dr. M. Hess, Lic. phil. S. Clematide, Lic. phil. G. Schneider

UEBERSICHT: CORPORA MIT LINGUISTISCHEN ANNOTATIONEN

NAME	**CORPUS**	**ANNOTATIONEN**	**VERFÜG-BARKEIT**	**GRÖSSE**	**CORPUS-SPRACHE**	**SPRACHE**	**LIT.HINWEIS**

ALEMBIC WORKBENCH	Schriftlicher Text	- ein Annotationssystem
		- Annotationen:
		  - gewöhnliche Textannotationen
		  - verschiedene Arten spezialisierter Annotationen inkl. Koreferenz-Annotationen
		  - benutzerdefinierte „inter-tag pointers“
		- „general template“ Annotationen (aka relations, frames, events)
			Frei verfügbar	Nicht angegeben	Multilingual:
					- Englisch
					- Spanisch
					- Japanisch
					- Chinesisch
					- Russisch
					- Griechisch
					- Thailändisch
						SGML	David Day: Mixed-Initiative Development of Language Processing Systems. 1997.

ANNOTATOR	Gesprochenes	- manuelle oder automatische Annotationen	Nicht verfügbar	Nicht angegeben	- Englisch - Japanisch - Spanisch - Cyrillic ?	Nicht angegeben	Nicht angegeben
Archivage	Spontangespräche	- phonologische Transkriptionen - freie Übersetzungen - Wort-zu-Wort Übersetzungen - Ton-/Klang-Annotationen für Dokumente auf dem Web	Originalaufnahmen nicht verfügbar	Nicht angegeben	New Caledonia-Sprachen mit Übersetzungen ins Französische	XML / XSL	Nicht angegeben
BNC (The Britisch National Corpus)	Geschriebenes (90%) und Gesprochenes (10%)	- AutomatischePart-of-speech (Wortklasse) Annotationen - Lemma Annotationen - Texteinteilung in orthographische Satzeinheiten - Strukturelle Texteinteilung (headings, paragraphs, lists usw.) - Vollständige Klassifikation jedes Textes mit kontextuellen und bibliographischen Informationen - Metatextuelle Informationen über die Ressource oder Entcodierung von individuellen Texten - Sprecherwechsel, Pausen, paralinguistische Merkmale in gesprochenen Texten	Verfügbar gegen Entgelt	Über 100 Millionen Wörter (1,5 Gigabytes)	Monolingual: Modernes Britisches Englisch	SGML	http://www.info.ox.ac.uk/bnc/news/index.html

BRAW reference corpus (Berlin-Brandenburgische Akademie der Wissenschaften)		- Morphosyntaktische Annotationen					
CA (Conversational Analysis)	Gesprächsaufnahmen	- Gesprächstranskriptionen - Prosodie abhörbar über einen Soundplayer Textsegmente annotiert mit Elementen der Prosodie	Frei verfügbar	Nicht angegeben	Englisch	Nicht angegeben	Emanuel A. Schegloff: „Reflections on Studying Prosody in Talk-In-Interaction“. 1997. → Paper inkl. Corpus
CES (The Corpus Encoding Standard)						SGML	
CHILDES (The Child Language Data Exchange System)	- Erst- und Zweitspracherwerbsdaten - Monolinguale und bilinguale Corpora	- Gesprächsaufnahmen - Transkriptionen - Vollständige morphologische und part-of-speech Analyse bei englischen Corpora - Verbindungen der Transkriptionen zu digitalem Tonmaterial und Bandaufnahmen („alignment of text to speech and video“)	Frei verfügbar	Auflistung der einzelnen Corpora mit den Grössenangaben in dieser Übersicht nicht möglich.	Über 30 Sprachen: - Englisch - Germanische und Nordische Sprachen - Romanische Sprachen - Andere Sprachen		Nicht angegeben
COCOSDA (The International Committee for the Coordination and Standardisation of Speech Databases and Assesment Techniques for Speech Input/ Output)	Gesprochenes						
CSAE (Corpus of Spoken American English)	Gesprächsaufnahmen	- Transkriptionen annotiert gemäss Transkriptionskonventionen (z.B.: Sprecherangaben, Pausen, Intonation, Dauer, Dialekte usw.) - Auf CD-ROM Transkriptionen mit Sound	Nur für Mitglieder der University of California, Santa Barbara	Nicht angegeben	Amerikanisches Englisch	Nicht angegeben	Nicht angegeben

CSLU (The Center for Spoken Language Understanding)	Telefongespräche	- Transkribierte Telefongespräche - Zurzeit erhältliche Corpora: - 22 Language - Alphadigt - Apple Words and Phrases - Cellular Words and Phrases - Foreign Accented English - ISOLET - Multi-Language Telephone Speech - Names - National Cellular - Numbers - Portland Cellular - Speaker Recognition - Spelled and Spoken Words - SR4X - Stories - Yes/No	Korpora gebührenfrei für Universitäten und Non-Profit-Organisationen	Nicht angegeben	Multilingual: Über 20 Sprachen	Nicht angegeben	http://cslu.cse.ogi.edu/toolkit/pubs/index.html
ELSNET/SSA (European Network in Speech and Natural Language Processing)		- Syntaktische und semantische Annotationen			Bilingual: - Deutsch - Italienisch		
FSA	Schriftlicher Text	- LADL (Laboratoire d'automatique Documentaire et Linguistique): Umfangreiche linguistische Datenbanken; elektronische Wörterbücher und Grammatiken					
GATE (General Architecture for Text Engineering) -> siehe Multext	Textsammlung	- Merkmale von GATE: - ein manueller Annotationstool - ein Annotationsvergleichstool - enthält drei Subsysteme: - GDM (GATE Document Manager) - GGI (GATE Graphical Interface) - CREOLE (Collection of REusable Objects for Language Engineering) - CREOLE übernimmt folgende Aufgaben: - Zerlegung in Tokens („tokenisation") - Identifizierung von Sätzen und Paragraphen - Part-of-speech Tagging	Für Forschungszwecke frei verfügbar			SGML	Michael Dorna: General Architecture for Text Engineering (GATE).1997.
Gsearch	Schriftliches und Mündliches	- ein Tool, um getaggte Corpora zu suchen arbeitet mit BNC, Brown, SUSANNE, WSJ, NEGRA und der Frankfurter Rundschau					

INTEX	Zeitungstexte	- ist ein Sprachverarbeitungssystem - enthält umfangreiche Wörterbücher und morphologische und syntaktische Grammatiken - annotiert umfangreiche Texte mit einem Index und baut lemmatisierte Konkordanzen („lemmatized concordances“) auf			Multilingual: - Englisch - Französisch - Deutsch - Griechisch - Italienisch - Polnisch - Portugiesisch - Spanisch - Bulgarisch		
ISIP (Institut for Signal and Information Processing)	Telephongespräche	- Segmentierung, Transkribierung und Annotationen von Telefongesprächen					
MATE (Multilevel Annotation, Tools Engineering)	Gesprächsdialoge	Annotationen (manuelle und halbautomatische): - Prosodie - Morpho-Syntax - Koreferenz - Dialogakte - Kommunikationsschwierigkeiten/-aspekte („communicative difficulties“) - Stufeninteraktion („inter-level interaction“)	Corpora nur für Mitglieder erhältlich			XML	http://mate.nis.sdu.dk/information/matepubl.html
Multext (Multilingual Text Tools and Corpora)	Zeitungstext	- Part of Speech Tagging für: - Englisch - Französisch - Deutsch - Spanisch - Italienisch - Annotationen: - Erkennung von Token und Sätzen - morpho-syntaktisches Tagging - Paralleltext-Gruppierung („alignment) - Prosodie-Markierung	Frei verfügbar		Multilingual: - Bambara - Bulgarisch - Catalanisch - Tschechisch - Niederländisch/Holländisch - Englisch - Estonisch - Französisch - Deutsch - Ungarisch - Italienisch - Kikongo - Occitanisch - Romanisch - Slovenisch - Spanisch - Swedisch - Swahili	SGML	

NEGRA (Nebenläufige grammatische Verarbeitung)	Zeitungstexte (Frankfurter Rundschau)	- Die Texte sind der CD „Multilingual Corpus I“ der European Corpus Initiative entnommen - Korpus mit Parts-of-Speech Annotationen versehen und vollständig mit syntaktischen Strukturen annotiert (inkl. Knoten und Kanten) - 10'000 Sätze sind linguistisch annotiert mit: - word (wordform) - cat (syntactic category) - pos (part-of-speech tags) - morph (morphological tag) - labeled edges - Die folgenden verschiedenen Typen von Informationen sind im Korpus kodiert: - Part-of-Speech Tags } terminale - Morphologische Analyse } Knoten - Grammatische Funktionen (Kanten) - Phrasale Kategorien (nichtterminale Knoten) - Das Korpus wurde semi-automatisch erstellt.	10‘000 Sätze (176'000 Tokens) sind frei verfügbar	350‘000 Tokens (ca. 20‘000 Sätze)	Deutsch	- intern als SQL-Datenbank - extern im zeilenbasierten Export-Format	Thorsten Brants: - The NeGra Export Format for Annotated Corpora. 1997. - An Annotation Scheme for free Word Order Languages. 1997.
Penn Treebank	Zeitungstexte	- ein umfangreiches und syntaktisch annotiertes Corpus - Aufgebaut in 2 Phasen: - 1. Phase des Penn Treebank Projekts: Vollständige POS-Annotationen und partielle „sceletal syntactic structure“ Annotationen - 2. Phase des Penn Treebank Projekts: Annotation von einfachen Prädikat-Argument Strukturen - vereinfachtes POS-Tagset; Tagset basiert auf demjenigen des Brown Corpus - halbautomatisch erstellte Syntaxstruktur-Annotationen	Begrenzte Verfügbarkeit	1‘200'000 Tokens (über 4,5 Millionen Wörter)	Amerikanisches Englisch		
Praat	Gesprächsdaten	- enthält Tools für Transkibierung und Annotierung auf mehrfachen Stufen					
SLAM (Segmentation and Labelling Automatic Module)	Gesprochenes	- Tool für halbautomatische Segmentierung und Etikettierung von Gesprächssignalen („speech signals“)					

SUSANNE (Surface and underlying structural analysis of natural Englisch)	Geschriebenes	- ein verständliches, vollständig explizites und detailliertes Annotationsschema für englische Grammatikstrukturen - Subset des „million-word“ Brown Corpus - Annotationen: - grammatische Kategorien (POS) - phrasale Kategorien - Repräsentation internaler Strukturen (Daten, Geldbeträge, Gewicht- und Mass-angaben, Multi-Wort-Personalnamen) - Keine expliziten Informationen über „head/modifier“ Beziehungen bei grammatischen Konstruktionen	Frei verfügbar	150'000 Tokens (130'000 Wörter)	Englisch		
TIGER Project	Zeitungstexte	- Aufbau eines linguistisch interpretierten, d.h. syntaktisch annotieren, Korpus deutscher Zeitungstexte - Extension des NEGRA Korpus in Grösse als auch im Detail der Annotationen - Jeder Satz im Korpus wurde halbautomatisch annotiert - Am Ende des Projekts wird der TIGER-Korpus über 50‘000 Sätze enthalten	Demo: Die ersten 250 Sätze des NeGra-Korpus	- Erster Schritt wurde mit dem NEGRA-Korpus gemacht; Umfang: 20'000 Sätze - Ziel: 50000 Sätze	Deutsch	XML	http://www.ims.uni-stuttgart.de/projekte/TIGER/paper.shtml
TransTool	Gesprochenes	Enthält folgende Tools: - TransTool (hilft beim Transkribieren) - Synchtool (zur Synchronisation von Transkriptionen mit Audio- und Videodaten) - TRASA (Tool für automatische Analyse des Corpus) - TRACTOR (Tool, um die Kodierung zu unterstützen)			Swedisch		
Verbmobil	Spontansprachliche Dialoge	- ein umfangreiches deutsches „speech-to-speech“ Übersetzungs-projekt - Annotationen: - Ortographie - teilweise Prosodieannotation (bei Übersetzungen des Deutschen ins Britische Englisch) - morphologisches und POS Tagging - Semantik-Annotation - Dialogakt-Annotation		Sprachpaar: - Deutsch-Englisch (ca. 10'000 Wörter) - Deutsch-Japanisch (ca. 2'500 Wörter)	Multilingual: - Deutsch - Englisch - Japanisch		http://www.ims.uni-stuttgart.de/projekte/verbmobil/vm-reports.html