Das Korpus
(Korpus)

0. Das Material für diese Analyse wurde größtenteils aus dem Gralis-Korpus übernommen. Das Gralis-Korpus stellt einen online abrufbaren, informationellen und analytischen Komplex für die Sammlung, Bearbeitung und Auswertung textueller, gesprochener und visueller Informationen zur systematischen Untersuchung slawischer Sprachen dar.[1] Es ist eine mehrsprachige, mehrdimensionale und multifunktionale Sammlung von Texten, Audio-, Video, TV- und anderen Aufnahmen, die für linguistische Untersuchungen zu slawischen Sprachen zusammengetragen und aufbereitet wurden.[2]

Der Zweck des BKS-Korpus besteht darin, in einer möglichst tiefen und umfassenden Untersuchung die Übereinstimmungen, Ähnlichkeiten und Unterschiede zwischen dem Bosni(aki)schen, Kroatischen und Serbischen zu eruieren. Angesichts dessen, dass es sich um nahe verwandte Sprachen handelt, deren Beziehung zueinander Grund für unterschiedliche Spekulationen liefert, soll mit der Erstellung eines solchen Korpus eine repräsentative und heterogene Quelle für eine objektive Beurteilung der Übereinstimmungen, Ähnlichkeiten und Unterschieden zwischen diesen Sprachen geschaffen werden. Weiters soll mithilfe des Korpus umfassendes Material für das Verfassen 1) eines komplexen korrelativen Wörterbuches der Sprachen Bs, Hr und Sr in einer gedruckten und einer Online-Version, 2) korrelativer Grammatiken des Bs, Hr und Sr und schließlich 3) eines Lehrbuchs des Bs, Hr und Sr zusammengetragen, aufbereitet und ausgewertet werden.

Das Gralis BKS-Korpus soll zeigen, wie sich die BKS-Einheiten (phonetisch-phonologische, orthoepische, grammatikalische und stilistische) auf sämtlichen Ebenen und auf Basis konkreten Materials in natürlicher Umgebung darstellen.

1. Das Gralis BKS-Korpus besteht aus zwei Korpora: dem Text- und dem Speech-Korpus. Beim Text-Korpus handelt es sich um eine Online-Sammlung paralleler Texte für verschiedene slawische Sprachen. Einen Teil des Gralis-Korpus bildet das Speech-Korpus. Dieses ist eine Online-Sammlung

---

[1] Der Name „Gralis“ leitet sich vom gleichnamigen, am 1. März 2000 eröffneten slawistischen Online-Portal der Karl-Franzens-Universität Graz her (http://www-gewi.kfunigraz.ac.at/gralis), wobei das Akronym Gralis für **Gra**zer **li**nguistische **S**lawistik steht.

[2] Die Vorstellung des Gralis-Korpus ist eine kurze Auswahl aus dem Sonderduck „Gralis-Korpus“ (Tošović 2008).

von Audiomaterial, das aus drei Subkorpora – dem Wort-, Fix- und Frei-Korpus besteht und über die Applikationen Akzentarium, Transkriptarium und Suprasegmentarium verfügt. Das Korpus-Interface kann in den Sprachen Bosni(aki)sch, Kroatisch und Serbisch abgerufen werden.

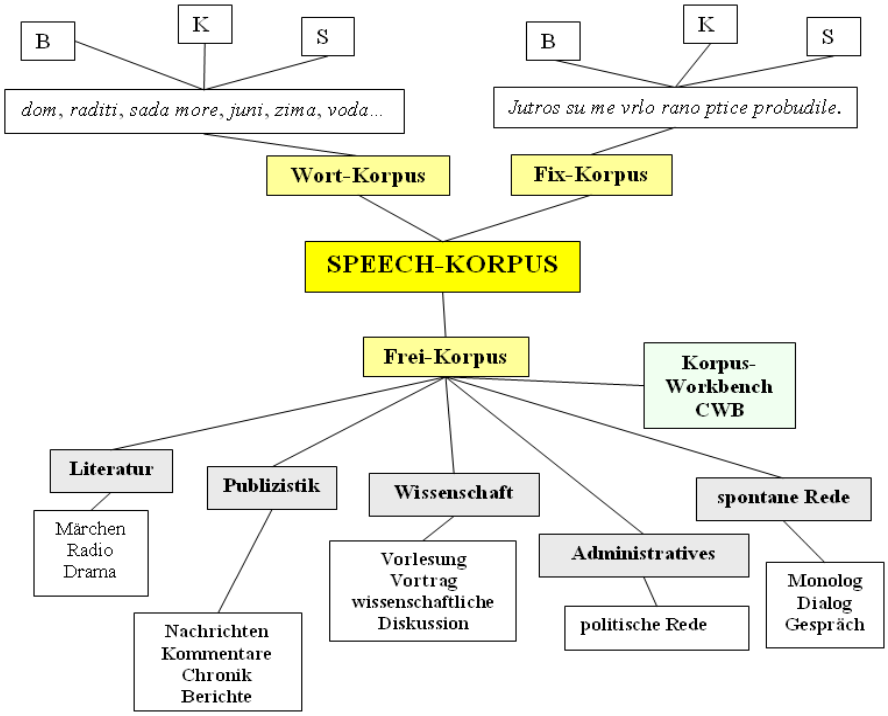

Abb. 2: Die Struktur des Gralis Speech-Korpus

2. Das W o r t - K o r p u s besteht aus Aufnahmen verlesener Wortlisten, und beim Fix-Korpus handelt es sich um Aufnahmen kürzerer Texte (der häufig aufgenommene Text „Jutro“ umfasst 18 Sätze; Tošović 2006c).

3. Das dritte Subkorpus im Rahmen des Gralis Speech-Korpus bildet schließlich das F r e i - K o r p u s , das zur Untersuchung spontan gesprochener Sprache dient. Angesichts der Tatsache, dass für ein solches Korpus keine vergleichbaren Beispiele bestehen (jede sprachliche Äußerung stellt ein Unikat dar und kann über kein semantisches Äquivalent verfügen), müssen Aufnahmen zu vergleichbaren Situationen (z. B. ein Gespräch am Markt, im Restaurant u. Ä.) oder Genres (Dialog, Erzählung, Diskussion, Entgegnung) getätigt werden. Dieses Subkorpus wird außerhalb der Struktur des auf einer MySQL-Datenbank basierenden Speech-Korpus entwickelt und fungiert als Teil des Text-Korpus, dem die Korpussoftware CWB zu Grunde liegt. Gegenwärtig umfasst das Frei-Korpus einzig eine Lebensschilderung, die im Buch „Ujak“ („Onkel“; Tošović 2003) abgedruckt wurde.

4. Von besonderer Bedeutung für die Bestimmung von Akzenten erweist sich das A k z e n t a r i u m , in dem durch Eingabe eines Suchbegriffes die jeweiligen Akzentuierungen in den Sprachen Bosni(aki)sch, Kroatisch, Serbisch und Serbokroatisch[3] angezeigt werden. Als Quelle für die einzelnen Lexeme dienen dabei Wörterbücher der bosnischen, kroatischen und serbischen Sprache. Mithilfe des Akzentariums wird die Akzentuierung von Wörtern in erheblichem Maße vereinfacht, indem man auf einen Blick die standardologischen Lösungen in Wörterbüchern der jeweiligen Sprachen angezeigt bekommt.

Im Zuge des Arbeitsschrittes des Eintragens der Akzente erfolgt zuerst die Wahl der Wörter mit den entsprechenden Akzenten, wobei das Programm den kanonischen Akzent (derjenige, der in lexikographischen Werken verzeichnet ist) als (in der Mehrzahl der Fälle) wahrscheinlichste Akzentuierungsvariante vorschlägt. Die graphische Darstellung entspricht dabei den klassischen, in der Orthoepie üblichen Symbolen.

Das Gralis-Akzentarium bietet die Option der Suche eines Akzentes bzw. von mehreren Akzenten im gesamten Speech-Korpus, wobei auch die Wahl einer Sprache (Bosni/aki/sch, Kroatisch, Serbisch) und einer konkreten Quelle vorgenommen werden kann. Im unteren Teil der Maske werden sodann all jene Lexeme aus dem Speech-Korpus angeführt, die mit dem entsprechenden Buchstaben beginnen.

Neben dem oben dargestellten Interface wurde für die Suche eine weitere Benutzeroberfläche entwickelt, in deren Mitte sich ein Fenster für den Eintrag eines gesuchten Lexems befindet.

Das Akzentarium beinhaltet ausschließlich Material aus Quellen, für die eine schriftliche Einverständniserklärung seitens der TrägerInnen der Urheberrechte vorliegt, wobei die Information zur Quelle durch einen Klick auf deren Abkürzung erscheint (z. B. Matešić 1966).

Ein bestimmtes Lexem kann auch nur innerhalb einer einzigen Sprache und innerhalb einer einzigen Quelle gesucht werden, wie etwa im „Mali akcenatski rečnik“ der serbischen Sprache von Milorad Dešić (2001).

Das Gralis Speech-Korpus bietet die Möglichkeit einer Akzentuierung der Audioaufnahmen, indem mittels Audio- oder Spektralanalyse die Akzente der einzelnen Lexeme festgelegt werden. Nach Durchführung einer auf Gehör basierenden Akzentuierung werden die Dateien unmittelbar nach Abschluss der Arbeiten ins Valorisarium überführt, in dem Fachleute eine Beurteilung der

[3] Im Falle von Publikationen, die zum Zeitpunkt des Bestehens der serbokroatischen Sprache erschienen sind.

Akzente vornehmen. Stimmen die Meinungen dreier ExpertInnen für Akzentologie überein, wird das Wort bzw. eine komplett bearbeitete Datei für die weitere Analyse freigegeben. Es sei generell darauf hingewiesen, dass dieser Arbeitsschritt für alle Beteiligten ein oftmaliges Abhören des Audiomaterials erforderlich macht.

Die technische Vorgangsweise zur Niederschrift der Transkription entspricht jenen zur Bestimmung der Akzente und der Satzintonation. Als primäres Alphabet dient dazu die Gralis-Transkription (die in diesen Buch erklärt und verwendet wird), die eigens für die Bedürfnisse der Sprachen Bosni(aki)sch, Kroatisch und Serbisch entwickelt wurde und später in international übliche Transkriptionsalphabete wie SAMPA oder IPA überführt werden kann.

Auf die Aufbereitung und Bearbeitung der Aufnahmen folgt die Analyse der einzelnen Dateien. Dies kann mittels Audio- aber auch durch eine Spektralanalyse mithilfe des Programms „Praat“ und des Skripts „Prosogramm“ geschehen, wobei genanntes Skript für eine eingehende und detaillierte Untersuchung der Intonation und des Akzentes dient. Daneben kann auch noch eine Analyse der Transkription vorgenommen werden.

Die für das BKS in besonderem Maße interessante Untersuchung der Akzente lässt sich mit dem auf Praat basierenden Programmskript „Prosogramm“ durchführen. Dieses Skript, das daneben auch noch Analysen anderer Art (etwa zur Intonation von Syntagmen und Sätzen) ermöglicht, fußt auf einer Reihe von errechneten Parametern, um einer dem menschlichen Ohr entsprechenden Perzeption nahe zu kommen. Für eine akustische Analyse der im Gralis Speech-Korpus enthaltenen Audioaufnahmen ist es erforderlich, zuerst eine Annotierung der Dateien in Praat durchzuführen und daneben auch ein TextGrid mit beigefügter Transkription anzulegen. Daraufhin kommt es zum Start des Prosogramms, wobei jeder Satz unter dem gleichen Namen und mit ansteigender Nummerierung abzuspeichern ist (z. B. *abc001.wav* für die Audiodatei und *abc001.TextGrid* für das TextGrid mit Transkription). Mit einem abschließenden Betätigen des Installationsfolders werden die Ergebnisse der vom Prosogramm automatisch berechneten akustischen Parameter graphisch dargestellt.

Durch die Ausarbeitung der bereits beschriebenen Aufnahmeevidenz als integraler Bestandteil des Gralis Speech-Korpus bietet sich die Möglichkeit, die biographischen Hintergründe jeder einzelnen Person abzurufen, wobei die Anonymität der ProbandInnen gewährleistet ist. Auf Grund der Elastizität des Korpus ist es zu jedem Zeitpunkt möglich, bereits eingetragene Angaben abzuändern und zu löschen, wobei diese ständige Bearbeitungsoption auch auf jedes Audiosegment (Satz und Wort) des Gralis Speech-Korpus zutrifft. Die Analysemöglichkeiten des Audiomaterials umfassen Intonation, Transkription und Akzentuierung und werden durch die zahlreichen Funktionen des Programms Praat und des Skripts Prosogramm wesentlich ausgeweitet.

5. Das am Institute of Phonetic Sciences an der Universität Amsterdam entwickelte Open-Source-Programm Programm Praat (dt. Übersetzung: „sprechen“) dient zur akustischen Analyse von Audiomaterial im Format wav. Praat (Praat-www) kann in verschiedensten Betriebssystemen (Windows, Linux u. a.) betrieben werden und ermöglicht ein breites Spektrum an phonetischen Analysen, die die Intensität, Intonation, Frequenz, Dauer, Formanten und andere artikulatorische Synthesen umfassen. Daneben können auch Segmentierungen und eine phonetische Transkription vorgenommen werden. Ein Spektrogramm stellt eine zeitlich-spektrale Darstellung des Tonverlaufs dar, wobei auf der Abszisse die Frequenz und auf der Ordinate die Amplitude abgebildet werden. Die im Rahmen des Gralis Speech-Korpus auf Satz- oder Wortebene segmentierten Texte können mithilfe von Praat bis auf die Phonemebene analysiert werden.

Das Gralis Speech-Korpus verfügt über mehrere Aufnahmequellen, bei denen digitale Diktiergeräte mit unterschiedlichen Aufnahmeoptionen zum Einsatz kommen. Dazu kommt auch die Nutzung verschiedener DVB-S-Empfänger und des Internets, um verschiedene TV- und Radiosendungen digital aufnehmen zu können. Nach der Aufnahme gilt es, das gesammelte Material zu bearbeiten, zu säubern und abschließend zu segmentieren, wobei für verschiedene Zwecke unterschiedliche Kriterien und Ansprüche vorzusehen sind. Für eine Sprachanalyse etwa wird ein verlustfreier Codec (Programm zur digitalen Kodierung und Dekodierung von Daten oder Signalen) benötigt, wogegen für eine Veröffentlichung im Internet ein Codec mit möglichst hoher Kompressionsrate erforderlich ist.