Branko Tošović, Arno Wonisch (ur.)

Srpski pogledi na odnose između srpskog, hrvatskog i bošnjačkog jezika

Die serbische Sichtweise des Verhältnisses zwischen dem Serbischen, Kroatischen und Bosniakischen

I/2

Institut für Slawistik der Karl-Franzens-Universität Graz
Beogradska knjiga
2010

Branko Tošović (Graz)

Das Gralis-Korpus

Im ersten Teil des Textblockes zum Gralis-Korpus werden dessen grundlegende Konzeption, Entstehung, die weiteren Entwicklungsrichtungen und dessen integrale Bestandteile vorgestellt. Der zweite Teil hat das Text-Korpus und das Speech-Korpus zum Thema. Im dritten Teil werden die technische Entwicklung des Korpus, die Arbeitsschritte des Aufnehmens, des Dekodierens und Bearbeitens von Sprachaufnahmen präsentiert. Der vierte Teil ist Programmen für eine automatische Segmentierung und Analyse von Audio- und Video-Aufnahmen (Gralis Audio-VideoTools), der Sammlung von Material mittels Online-Umfragen (Gralis-Anketarium) und der Online-Begutachtung (Gralis-Rezensarium) gewidmet. Im abschließenden fünften Teil folgen Beiträge zu unterschiedlichen Programmen, wie etwa zur Sammlung und Verwaltung von bilbliographischen Einheiten slawischer Sprachen (Gralis-Bibliothekarium), zur Administrierung personenbezogener Angaben über die an Projekten mitarbeitenden Personen (Gralis-Personalium) und zu einem Programm für das Studium intersprachlicher orthographischer Korrelationen (Gralis-Präskriptarium).

1. Zum Studium slawischer Sprachen ist es überaus wichtig, über komplexes und in funktional-stilistischer Hinsicht ausgewogenes Material zu verfügen, auf das online zugegriffen werden kann. Dies trifft umso mehr auf komparative Untersuchungen nahe verwandter slawischer Sprachen, wie etwa im Falle von bosnisch/bosniakisch, kroatisch und serbisch (im Folgenden: BKS, B, K, S oder B/K/S) zu. Für derartige Analysen können zwei Arten von elektronischen Korpora herangezogen werden: Einerseits monolinguale Korpora, die zum Studium einer einzigen Sprache ohne Vergleichsmöglichkeiten mit anderen Sprachen vorgesehen sind. Derartige Korpora gibt es für beinahe alle slawischen Sprachen (Das Nationalkorpus der russischen Sprache – Национальный корпус русского языка, Das Nationalkorpus der russischen Literatursprache – Национальный корпус русского литературного языка – Narusco, Das Internetkorpus der weißrussischen Sprache – Интернет-корпус белорусского языка, Das tschechische Nationalkorpus – Český národní korpus, Das slowakische Nationalkorpus – Slovenský národný korpus, Das Korpus des Institutes für Informatik der Polnischen Akademie der Wissenschaften – Korpus Instytuta Podstaw Informatyki Polskiej Akademii Nauk – IPI PAN, Das Korpus der slowenischen Sprache FIDAPlus – Korpus slovenskega jezika FIDAPlus, Das Korpus gesprochener slowenischer Sprache – Korpus govorjene slovenščine, Das Korpus gesprochener bulgarischer Sprache – Корпус от разговорен български език u. a.). Im Falle des B/K/S kann auf zwei kroatische Korpora (Das kroatische Nationalkorpus – Hrvatski nacionalni koprus, Kroatische „Online-Schatzkammer“ – Hrvatska mrežna riznica) und ein serbisches Korpus (Korpus der modernen serbischen Sprache an der Mathematischen Fakultät der Universität Belgrad – Korpus savremenog srpskog jezika na Matematičkom

fakultetu Univerziteta u Beogradu) zurückgegriffen werden.[1] Daneben gibt es auch ein kleineres Korpus bosnischer Texte an der Universität Oslo, das jedoch gegenwärtig leider nicht zugänglich ist. Die zweite Art von Korpora bilden parallele (bi- oder polylinguale) Korpora, die für Untersuchungen von zumindest zwei Sprachen herangezogen werden können. Beispiele dafür lassen sich in der Slawia leider kaum antreffen, wodurch die Möglichkeit komparativer, kontrastiver oder korrelationaler Analysen slawischer Sprachen kaum gegeben ist. Ein diesbezüglicher Bedarf ist ohne Zweifel vor allem bei Analysen zu sehr nahe verwandten Sprachen (wie eben des BKS) anzutreffen, um innerhalb eines Kontextes und im direkten Kontakt textueller Einheiten die Übereinstimmungen, Ähnlichkeiten und Unterschiede wie auch Nuancen in Bedeutung und Gebrauch erfassen zu können. Angesichts des Fehlens eines solchen Korpus wurde deshalb der Versuch unternommen, im Rahmen des vorliegenden FWF-Projektes ein trilinguales Korpus für das B, K, S zu entwickeln, das mit seinen beiden Subkorpora – Text-Korpus und Speech-Korpus – sowohl textuelle als auch auditive Analysen ermöglicht. Auf Grundlage dieses BKS-Korpus wurden in weiterer Folge die Konzeption und Infrastruktur für die Erstellung von Parallelkorpora für andere slawische Sprachen geschaffen, die den gemeinsamen Namen Gralis-Korpus tragen. Eine wesentliche Komponente dieses Korpus liegt auch darin, dass slawische Sprachen direkt mit dem Deutschen verglichen werden können.

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. 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. Das Gralis-Portal befindet sich auf einem Server der Geisteswissenschaftlichen Fakultät (www-gewi.uni-graz.at) der Karl-Franzens-Universität Graz und nimmt 55 Prozent des Serverspaces ein.[2] Gegenwärtig setzt sich Gralis aus über 3.000 Websites zusammen, die folgende integrale Teile des Portals umfassen: Projektarium, Korpusarium, Educarium, Gralisarium, Grazer Slawisticarium und Operarium.

[1] Ein weiteres Korpus – das Korpus der serbischen Sprache von Đorđe Kostić (Корпус српског језика Ђорђа Костића) – ist nicht online zugänglich.

[2] Laut Angaben von Herrn Dieter Schicker (Serveradministrator am Institut für Informationsverarbeitung in den Geisteswissenschaften – INIG) vom 11.10.2007 stellt sich das Verhältnis Gralis-Portal vs. andere Portale, Anwendungen u. Ä. der Geisteswissenschaftlichen Fakultät wie folgt dar: von 20 GB werden 11 von Gralis und die restlichen 9 von anderen BenutzerInnen der Fakultät belegt. Dies erklärt sich dadurch, dass das Portal eine große Zahl an sehr viel Speicherplatz einnehmenden Audio-und Videodateien enthält.

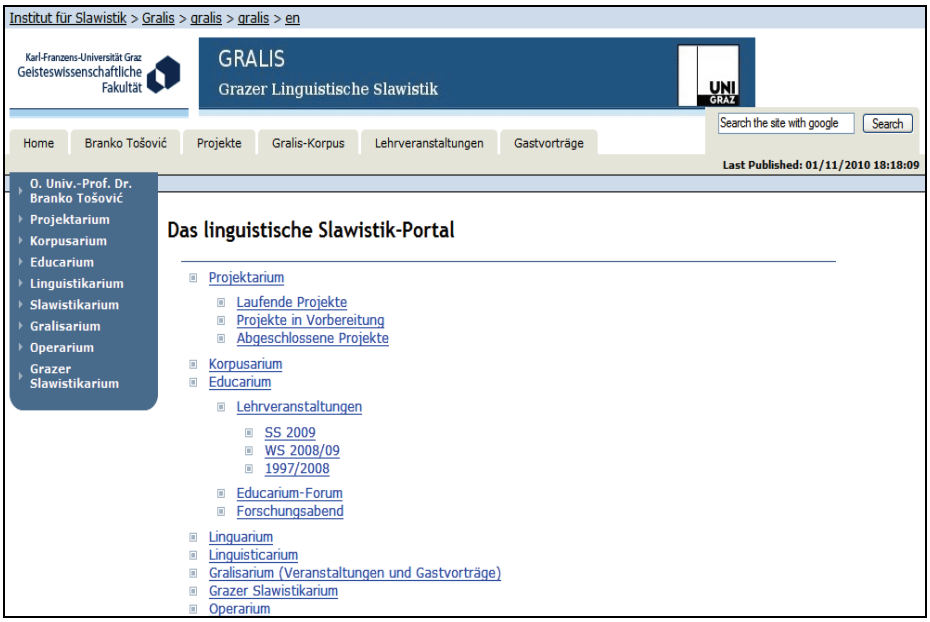


Abb. 1: Die Gralis-Startseite

Das Projektarium bildet eine Plattform zur Sammlung, Bearbeitung und Analyse linguistischen Materials im Rahmen von Forschungsprojekten wie (1) „Die Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen“ [„Razlike između bosanskog/bošnjačkog, hrvatskog i srpskog jezika“] (FWF-Projekt, P19158-G03, 2006–2009), (2) „Die vergleichende Analyse der semantisch-derivativen Kategorie der Aktionsarten in der slawischen Sprachen“ [„Studium porównawcze nad kategorią semantyczno-słowotwórczą Aktionsarten w językach słowiańskich“] (Projekt des Ministeriums für Wissenschaft und höheres Schulwesen der Republik Polen, Nr. N104 012 31/0898, 2006–2009) u. a. Weiters dient das Projektarium für öffentliche und frei zugängliche Präsentationen slawischer Sprachwissenschaftsprojekte und darüber hinaus auch als Medium für all jene, die einen Beitrag zu wissenschaftlichen Untersuchungen slawischer Sprachen leisten möchten. Als besonderer Schwerpunkt im Rahmen der Rubrik Projektarium wurde im Herbst 2005 eine einmal im Monat am Institut für Slawistik stattfindende Veranstaltungsreihe mit dem Titel „Forschungsabend“ (www-gewi.kfunigraz.ac.at/gralis/6.Educarium/Forschungsabend/Forschungsabend. htm) initiiert, die Studierenden bei der Abfassung wissenschaftlicher Arbeiten behilflich sein und generell zur Förderung der wissenschaftlichen Betätigung von Studierenden dienen soll.

Das Educarium stellt eine Online-Plattform für das Erlernen slawischer Sprachen dar, die sich aus dem Grazer Educarium, dem Linguarium und dem Linguisti-

carium zusammensetzt. Das Grazer Educarium beinhaltet Material für den Unterricht zu Disziplinen der slawischen Sprachwissenschaft und besteht aus vier Teilen: Der erste betrifft den Unterricht auf dem Gebiet der slawischen Linguistik am Institut für Slawistik der Karl-Franzens-Universität Graz, der zweite trägt die Bezeichnung Educarium-Forum und dient als Hilfsmittel für den Unterricht sowie einen wechselseitigen Informationsaustausch zwischen Lehrenden und Studierenden. Der dritte Teil nennt sich BKS-Abend und ist Themen des Unterrichts der Sprachen bosnisch/bosniakisch, kroatisch und serbisch gewidmet, und im vierten Teil mit dem Titel Dissertarium werden schließlich Dissertationen, Diplom- und andere Arbeiten präsentiert und Informationen zu Diplomprüfungen weitergegeben. Besondere Teile des Grazer Educariums stellen das Textarium (Sammlung von für den Unterricht vorgesehenen Texten) und das Translatorium (mit elementaren, für Studierende der Slawistik vorgesehenen Informationen aus der Theorie und Praxis des Übersetzens und Dolmetschens) dar.

Das Linguarium bietet (in erster Linie Studierenden) grundlegende Informationen zu sämtlichen slawischen Sprachen und besteht aus folgenden Rubriken: Slawische Sprachen, Altkirchenslawisch, B/K/S (Bosnisch/Bosniakisch, Kroatisch, Serbisch, Montenegrinisch, Serbokroatisch), Bulgarisch, Burgenlandkroatisch, Kaschubisch, Mazedonisch/Makedonisch, Polnisch, Russisch, Rusinisch/Ruthenisch, Slowakisch, Slowenisch, Sorbisch, Tschechisch, Ukrainisch und Weißrussisch. Das Linguisticarium enthält Informationen zur Slawistik, Sprachwissenschaft und zu den wichtigsten linguistischen Disziplinen (Linguistik, Graphik, Orthographie, Phonetik, Phonologie, Grammatik, Morphologie, Syntax, Lexikologie, Lexikographie, Phraseologie, Wortbildung, Textgrammatik, Stilistik, Soziolinguistik, Dialektologie, Computerlinguistik).

Beim Grazer Slawistikarium handelt es sich um eine Plattform zur Präsentation der slawischen Sprachwissenschaft in Graz, die sich aus drei Teilen – Forschungstätigkeit, Forscher und Lehrtätigkeit – zusammensetzt. Im Rahmen der Forschungstätigkeit werden dabei folgende Aspekte der Grazer Slawistik dargestellt: Geschichte, Perspektiven, Forschungsprofil, untersuchte Sprachen, Projekte, Publikationen, Kooperation, wissenschaftliche Veranstaltungen, Dissertationen, Diplomarbeiten. Die Rubrik mit dem Titel „Ich bin ein/e GrazerIn" beinhaltet Informationen zu auf dem Institut für Slawistik in Graz abgehaltenen Lektoraten, Gastvorträgen, Kongressen usw. In der Unterrubrik mit der Bezeichnung Forscher werden in einer Gliederung nach drei Zeitabschnitten grundlegende Informationen zu am Grazer Institut für Slawistik tätigen ForscherInnen präsentiert. Es handelt sich dabei **(1)** um das 19. und 20. Jahrhundert (Gregor Krek, Karel Štrekelj, Vatroslav Oblak, Matija Murko, Fran Ramovš, Rajko Nahtigal, Heinrich Felix Schmid, Bernd von Arnim und Josef Matl), **(2)** um das 20. Jahrhundert (Linda Aitzetmüller-Sadnik, Stanislaus Hafner, Herbert Schelesniker, Harald Jaksche und Erich Prunč) und schließlich **(3)** um Personen, die sowohl im

20. als auch im 21. Jahrhundert an der Grazer Slawistik tätig waren bzw. sind (**a**: auf dem Gebiet der Sprachwissenschaft: Maximilian Hendler, Ludwig Karničar, Heinrich Pfandl, Branko Tošović und Manfred Trummer, **b**: in der Literatur-, Kultur- und Sprachwissenschaft: Wolfgang Eismann, Peter Grzybek sowie **c**: in der Sprachbeherrschung: LektorInnen, Lehrbeauftragte u. a.). Der letzte Teil des Grazer Slawistikariums beinhaltet Angaben zu sprachwissenschaftlichen Lehrveranstaltungen aus den drei Studienrichtungssprachen (BKS, Russisch und Slowenisch), aus den Lektoratssprachen (Bulgarisch, Polnisch, Tschechisch) und Allgemeines zu lebenden slawischen Sprachen sowie zu Altkirchenslawisch. Eine weitere Kategorisierung betrifft die Sprache der Lehrtätigkeit von am Institut tätigen Personen, wobei zwischen den Sprachen der primären und sekundären Lehrtätigkeit unterschieden wird.

Die Rubrik Gralisarium bietet (beginnend ab 1997) Informationen zu wissenschaftlichen Veranstaltungen und Gastvorträgen auf dem Institut für Slawistik der Karl-Franzens-Universität Graz.

Das Operarium setzt sich aus unterschiedlichsten Informationen für wissenschaftliche und edukative Aktivitäten zusammen und besteht aus den Unterpunkten Internetarium, Online-Wörterbücher, Formulare, GIS, ZID, Formulare des Personalwesens der Uni Graz, UNIGRAZonline, Webmail, Einladung von Gästen und Aktuelles.

Den nun abschließend beschriebenen Bestandteil von Gralis bildet das Koprusarium, das als Plattform für die Aufbereitung, Bearbeitung, Analyse und Online-Präsentation von Korpusmaterialien dient und dessen wesentlichsten Bestandteil das Gralis-Korpus darstellt. Daneben bietet das Korpusarium Informationen zu den wichtigsten Korpora im Rahmen der Slawia, zu Korpora anderer Sprachen (englisch, deutsch u. a.) und im Besonderen zu Fragen der Korpuslinguistik.

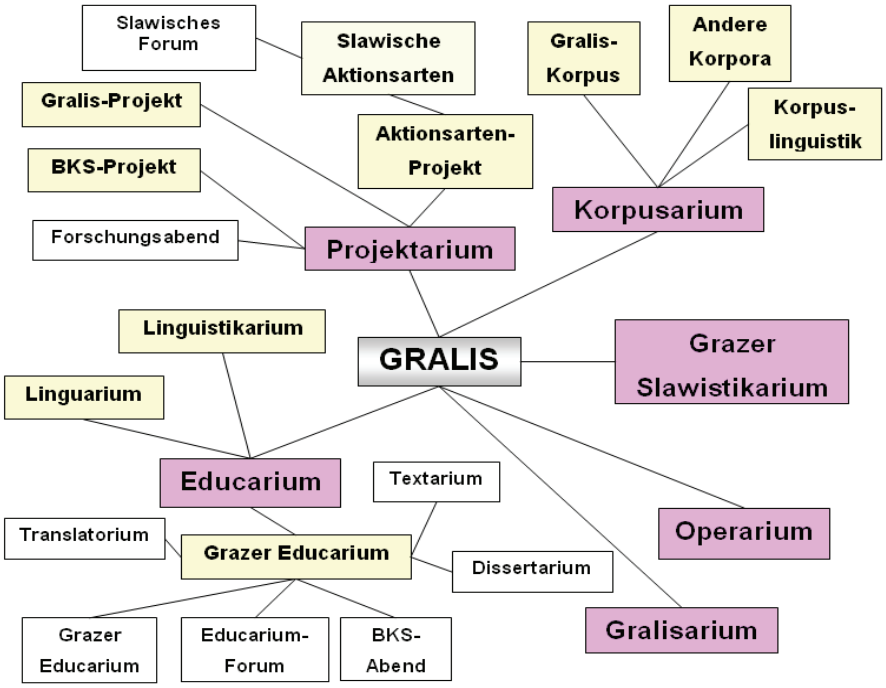


Abb. 2: Die Struktur von Gralis

2. Das Gralis-Korpus stellt eine online zugängliche, mehrsprachige, mehrdimensionale und multifunktionale Sammlung von Texten, Audio-, Video, TV- und anderen Aufnahmen dar, die für linguistische Untersuchungen zu slawischen Sprachen zusammengetragen und aufbereitet wurden. Es besteht aus drei großen Teilen, die mit den Bezeichnungen Gralis-Korporarium, Gralis-Komplementarium und Gralis-Tools versehen wurden.

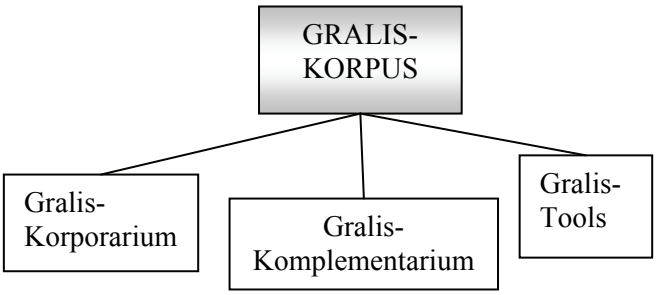


Abb. 3: Die Struktur des Gralis-Korpus

Mit der Entwicklung des Korpus wurde im Jahr 2006 begonnen, wobei sich das (seit diesem Zeitpunkt im Großen und Ganzen unveränderte) Korpusteam aus folgenden Personen zusammensetzt: dem Korpusleiter (Branko Tošović), dem Korpuskoordinator (Arno Wonisch), einer Person für die Erstellung relationaler Datenbanken im MySQL-Format (Olga Lehner, ab 2007), einer Person für die technische Leitung und Umsetzung, für die Textverarbeitung in den Formaten XML und TEI sowie für die serverfertige Adaptierung von Texten (Hubert Stigler, ab 2006), einem Administrator für die Schnittstellenprogrammierung (Dieter Schicker, ab 2006), einer Webdesignerin (Martina Semlak, ab 2007), einem Programmierer der Rezensariums (Stefan Kofler, ab 2007), einem Programmierer des Anketariums (Robert Thomann, ab 2007), einem für technische Unterstützung und die Gralis-Audio- und Video-Skripts verantwortlichen Mitarbeiter (Boris Tošović, 2006–2007) sowie mehreren MitarbeiterInnen für die Sammlung und Bearbeitung von Text-, Audio- und Videomaterial (Sandra Forić, ab 2006; Maja Midžić, ab 2006; Elvira Skledar, 2006; Alexander Just, 2006–2007 und Daniel Dugina, ab 2007). Bei der Erstellung des Korpus standen mit Vorschlägen, Hinweisen und Ratschlägen sowie in mehreren Beratungen Fachleute für die Korpuslinguistik aus Belgrad (Duško Vitas, Miloš Utvić, Cvetana Krsteva, Ranka Stanković und Ivan Obradović, 2006–2007), Chandler/Arizona (Danko Šipka, 2006–2007), Ljubljana (Tomaž Erjavec, 2006–2007), Moskau (Dmitrij Dobrovoljski, 2006), Zadar (Damir Ćavar, 2006), Zagreb (Marko Tadić, 2006) und Graz (Kurt Tiefenbacher, 2006) hilfreich zur Seite. An der Entwicklung des Gralis Speech-Korpus waren ExpertInnen aus Novi Sad (Milan Sečujski, 2007), Genf (Tea Pršir, ab 2007), Ljubljana (Jana Zemljarič-Miklavčič, 2006) und Moskau (Svetlana Savčuk, 2007) wesentlich beteiligt. Bei der Ausarbeitung des Akzentariums konnte auf die wertvollen Hinweise von Fachleuten aus Zagreb (Elenmari Pletikos, 2007 und des mittlerweile verstorbenen Ivan Ivas, 2006) zurückgegriffen werden. Bei der Bereitstellung von akzentuiertem Sprachmaterial waren bei der Erstellung des Akzentariums in hohem Maße Josip Matešić aus Mannheim (2007) und Milorad Dešić aus Belgrad (2007) behilflich. Die Überprüfung der von ProjektmitarbeiterInnen eingetragenen Akzente erfolgte durch Dragomir Kozorama aus Banjaluka (2007), Milan Tasić und Milorad Dešić aus Belgrad (2007). Von großer Bedeutung war die Übernahme umfangreichen Audiomaterials von Gesprächen mit den bekanntesten SlawistInnen des ehemaligen Jugoslawiens, die vom Publizisten Miloš Jevtić im Zweiten Programm des Belgrader Radios geführt und von diesem für das Frei-Korpus zur Verfügung gestellt wurden (2007).[3]

Bei der Entwicklung des Wort- und Fix-Korpus war in erheblichem Maße Rudolf Muhr aus dem Institut für Germanistik der Karl-Franzens-Univversität Graz

[3] Mehr dazu siehe im Beitrag von Miloš Jevtić in diesem Band.

beteiligt (ab 2007), der für die Erstellung dieser Korpora das von ihm entwickelte Programm Adaba zur Verfügung stellte. Bei der Planung und den ersten Arbeitsschritten für die Schaffung eines Spracherkennungsprogramms mit der Bezeichnung „BKS-Voice“ waren die Hinweise von Herrn Siegfried Kunzmann aus München (2006), Igor’ Chejdorov aus Minsk (2006–2007), Sanda Martinčić-Ipšić aus Rijeka (2006–2007), Vera Aleksić von der Firma Linguatec in München (ab 2006) wie auch von den Fachleuten von der Technischen Universität Graz, Gernot Kubin (ab 2006), Stefan Petrik (ab 2007) und Denis Helić (2006), von großer Hilfe.

Während einer Forschungsreise nach Zagreb (Kroatien), Belgrad (Serbien), Sarajevo und Mostar (Bosnien und Herzegowina) im von 13. bis 19. April 2006 wurde im Rahmen von Beratungen die Konzeption des Gralis-Korpus vorgestellt und gemeinsam mit den GesprächspartnerInnen analysiert. Ein weiterer dieser Forschungsaufenthalte des Korpusleiters führte im Februar 2007 nach, wo im Folgenden angeführte Konsultationen mit russischen Fachleuten auf dem Gebiet der Korpuslinguistik geführt wurden, die sich als überaus nützlich herausstellen sollten. Es waren dies in erster Linie Gespräche mit dem Leiter des Russischen Nationalkorpus, Vladimir Plugnjan (Institut für die russische Sprache „V.V. Vinogradov“ der Russischen Akademie der Wissenschaften), mit Angehörigen des EDV-Zweiges des genannten Institutes (Anatolij Šajkevič, Svetlana Savčuk u. a.), mit den Mitarbeitern des Institutes für theoretische und angewandte Sprachwissenschaft der Moskauer staatlichen Universität: Aleksandr Kibrik (Institutsleiter), Ol’ga Krivnova (Leiterin einer Gruppe zur Durchführung einer automatischen Synthese und Erkennung der russischen Sprache) und Sandro Kodzasov (Mitglied der genannten Gruppe).

Für die theoretische Konzeption und Vorbereitung des Gralis-Korpus erwies sich ein vom Korpusleiter im Sommersemester 2006 veranstaltetes Seminar von wesentlicher Bedeutung. Bei dieser Lehrveranstaltung waren folgende Fachleute auf dem Gebiet der Korpuslinguistik mit Vorträgen zu Gast: Damir Ćavar (erklärte die Konzeption und Struktur der Hrvatska mrežna riznica), Dimitrij Dobrovoljski (stellte das Russische Nationalkorpus vor), Tomaž Erjavec (demonstrierte das Korpus der slowenischen Sprache FIDAPlus und erläuterte das von ihm entwickelte Programm Multext-East), Bernhard Kettemann vom Institut für Anglistik der Karl-Franzens-Universität Graz (hielt ein Referat mit dem Thema „Korpus von Intelligent Design Texten“), Stefan Schneider vom Institut für Romanistik der Karl-Franzens-Universität Graz (zeigte das Online-Korpus BADIP – Banca dati dell'italiano parlato), Danko Šipka (hielt ein Referat zum Thema „Textkorpora in angewandter Slawistik“), Marko Tadić (sprach über das Kroatische Nationalkorpus) und Duško Vitas (präsentierte das Korpus der modernen serbischen Sprache an der Mathematischen Fakultät der Universität Belgrad).

Im Rahmen des Seminars kam es zur Präsentation der wichtigsten slawischen Korpora, elektronischen Bibliotheken und Wörterbücher, wobei von den genannten Studierenden folgende Themen vorgetragen wurden: Angloamerikanische Korpora (Gudrun Krenn), Bosnische und serbische digitale Bibliotheken (Goran Pajičić), das Bulgarische Nationalkorpus (Iva Hristova und Petya Dimitrova), das Tschechische und das Slowakische Nationalkorpus (Rita Plos und Corinna Schnedhuber), Deutsche einsprachige Textkorpora (Karin Markut), Einführung in die Korpuslinguistik (Branko Tošović), das Gralis-Korpus (Arno Wonisch), Was ist ein Korpus? (Branko Tošović), Korpus bosnischer Texte an der Universität Oslo (Maja Midžić und Sandra Forić), Korpus der serbischen Sprache von Đorđe Kostić (Marija Redi), Korpus des Institutes für Informatik der Polnischen Akademie der Wissenschaften (IPI PAN – Arno Wonisch), Kroatische Parallelkorpora (Silvije Beus und Ernedina Muminović), Kroatische Rohkorpora und digitale Bibliotheken (Elvira Skledar), Parallelkorpora (Florian Thelen), Russische Korpuslinguistik im Internet (Andreas Konrad und Doris Weißenböck), Slawische Korpuslinguistik (Branko Tošović und Arno Wonisch), Slawisch-französische Textkorpora (Ruth Aigner und Linde Prenn), Slawische Korpuslinguistik (Andreas Krammer und Theresa Križaj), Ukrainische und weißrussische Korpuslinguistik (Andreas Schiestl) sowie WordNet und RussNet (Tanja Eder).[4]

Die endgültige Ausgestaltung der Konzeption des Korpus erfolgte schließlich im Vorfeld des von 12. bis 14. April 2007 in Graz abgehaltenen 1. Projekt-Symposiums, das den phonetisch-phonologischen, orthoepischen und orthographischen Unterschieden zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen gewidmet war und dessen Programm auch eigene Themenblöcke namens Gralis-Korpus und BKS-Voice umfasste. Die in diesen Sektionen präsentierten Referate und Diskussionen von Vera Aleksić, Tomaž Erjavec, Igor' Chejdorov, Cvetana Krstev, Sanda Martinčić-Ipšić, Ivan Obradović, Ranka Stanković, Stefan Patrik, Svetlana Savčuk, Milana Sečujski, Hubert Stigler, Miloš Utvić und Duško Vitas brachten wesentliche Aspekte hinsichtlich der Sammlung und Bearbeitung von Korpustexten zum Vorschein. Auf diesem Symposium kam es schließlich auch zur offiziellen Eröffnung des Gralis-Korpus. Einen Monat später, am 31. Mai 2007, wurde das Korpus im Rahmen einer Informationsveranstaltung des Institutes für Informationsverarbeitung in den Geisteswissenschaften durch den Korpusleiter ein zweites Mal einer breiteren Öffentlichkeit vorgestellt.

Im Zuge der Vorarbeiten zur Entwicklung des Korpus wurden im Rahmen der Gralis-Aktivitäten 2006 auch einige weitere Veranstaltungen abgehalten, bei denen

[4] Die Nennung aller Korpora, Bibliotheken und Wörterbücher erfolgt entsprechend den Titeln der Referate in deutscher Sprache.

Cvetana Krstev (Referat zu elektronischen Wörterbüchern), Duško Vitas (automatische Textbearbeitung) und Jana Zemljarič-Miklavčič (Korpus der gesprochenen slowenischen Sprache) wertvolle Aspekte aufzuzeigen vermochten. Im Jahre 2007 wurden diese Aktivitäten mit Vorträgen von Milan Sečujski (Automatische morphologische Annotation im Lichte der Besonderheiten des BKS) und Stefan Petrik (Grundlagen der Spracherkennung) fortgesetzt.

Im September 2006 wurde von Miloš Utvić von der Mathematischen Fakultät der Universität Belgrad für alle am Projekt mitarbeitenden Personen ein sechstägiger Kurs mit dem Thema „Textverarbeitung, Etikettierung, Parallelisierung und Vertikalisierung bei der Erstellung von Korpora“ abgehalten.

Für die Entwicklung des Gralis Speech-Korpus erwiesen sich im Folgenden genannte, im Jahre 2007 abgehaltene Veranstaltungen als überaus hilfreich und nützlich: **(1)** die Vorträge von Rudolf Muhr zu Themen betreffend Korpora der gesprochenen Sprache – **a)** Zur Theorie der plurizentrischen Varietäten des Deutschen, **b)** Zur Phonetik der Varietäten des Deutschen, **(2)** die Ausführungen von Milan Tasić hinsichtlich der Ausarbeitung des Gralis-Suprasegmentariums (Intonation in der modernen serbischen Sprache), **(3)** das Referat von Milorad Dešić in Bezug auf das Gralis-Akzentarium (Der Akzent in der serbischen Standardsprache), **(4)** der Vortrag von Tea Pršir im Lichte der akustischen Bearbeitung von Audiomaterial (Vergleichende Prosodie des BKS mithilfe des Prosogramms), **(5)** die Darlegungen von Dragomir Kozomara zur Ausarbeitung der Gralis-Präskriptariums (Lexikalisch-orthographische Zweifelsfälle in der serbischen Sprache) und **(6)** die Präsentation von Vera Aleksić angesichts der Entwicklung von BKS-Voice (Sprachtechnologien und moderne Methoden der Spracherkennung). Ebenfalls im gleichen Jahr wurde den Studierenden des Institutes für Slawistik von den KorpusmitarbeiterInnen Sandra Forić, Olga Lehner, Maja Midžić und Arno Wonisch am 23. Mai 2007 erstmals das Gralis Speech-Korpus in seinem gesamten Umfang präsentiert. Informationen zu allen angeführten (Gast)vorträgen und Referaten stehen allen Interessierten in der Rubrik Gralisarium des Gralis-Portals zur Verfügung (http://www-gewi.kfunigraz.ac.at/gralis/4.Gralisarium/Gralisarium.htm).

Als Tribüne für unterschiedliche Fragen in Bezug auf die Entwicklung des Gralis-Korpus erwies sich der einmal monatlich durchgeführte Forschungsabend, der vor allem dazu dient, Studierenden Aspekte wissenschaftlicher Betätigung aufzuzeigen und ihnen Modelle und Nutzungsmöglichkeiten von Korpora nahe zu bringen. Angesichts dessen, dass ein Teil des Korpusmaterials durch relationale Datenbanken verwaltet wird, wurden von Dieter Schicker (Institut für Informationsverarbeitung in den Geisteswissenschaften – INIG) im Rahmen von vier Forschungsabenden (27. April, 3. Mai, 7. und 14. Juni 2006) kurze Kurse mit dem Titel „Einführung in SQL anhand der freien Datenbanksoftware MySQL“ abgehalten. Ein weiteres Resultat der Forschungsabende liegt darin, dass in

mehreren Diskussionen die Erkenntnis gewonnen wurde, dass im Rahmen des Sammelns von Quellen für wissenschaftliche Arbeiten eine Online-Befragung von großem Nutzen sein kann. Dies kam besonders deutlich beim am 14. Dezember 2006 abgehaltenen 11. Forschungsabend zum Ausdruck, bei dem Michaela Handke ein Referat mit dem Titel „Der Nutzen von Umfragen und Fragenbogen für studentische wissenschaftliche Arbeiten“ vortrug. Ab diesem Zeitpunkt wurde mit der Ausarbeitung des Gralis-Anketariums begonnen, das von Robert Thomann im Herbst 2007 erfolgreich fertig gestellt werden konnte und Studierenden erstmals beim 17. Forschungsabend am 21. November 2007 präsentiert wurde (Branko Tošović – Arno Wonisch: Erstellen von Online-Umfragen für Seminar- und Diplomarbeiten mithilfe des „Gralis-Anketariums“).

Im Rahmen des Forschungsabends wurden weiters auch Fragen der Spracherkennung (Stefan Petrik: Grundlagen der Spracherkennung, 14. Juni 2007), der akustischen Analyse (Tea Pršir: Vergleichende Prosodie des BKS mithilfe des Prosogramms, 7. Oktober 2007; Arno Wonisch – Sandra Forić: Nutzung akustischer Analysen slawischer Sprachen für studentische Arbeiten, 29. März 2007) und von Parallelkorpora (Arno Wonisch: Paralleltextkorpora, 30. November 2006) erörtert.

Im Laufe der Jahre 2006 und 2007 nahmen die am Korpus mitarbeitendenden Personen an mehreren Konferenzen und Tagungen teil und stellten dabei Aspekte des Gralis-Korpus vor. Es handelte sich dabei um Referate, in denen einerseits entweder das Korpus als **(1)** Hauptthema fungierte, wie etwa **(a)** bei der 21. Tagung der Kroatischen Gesellschaft für angewandte Linguistik mit dem Thema „Sprachpolitik und Sprachrealität“ (Branko Tošović – Arno Wonisch: Gralis-Korpus, Split /Kroatien/, Mai 2007), **(b)** auf der 12. Internationalen Slawistiktagung (Branko Tošović: Korporaaspekte der kroatisch-serbischen sprachlichen Berührungspunkte, Opatija /Kroatien/, Juni 2007), **(c)** bei der selben Tagung (Hubert Stigler – Arno Wonisch: Das Gralis-Korpus als Plattform zum Studium kroatisch-serbischer sprachlicher Berührungspunkte, Opatija, Juni 2007) und **(d)** auf der 6. Internationalen Tagung „Untrersuchungen zur gesprochenen Sprache“ (Daniel Dugina – Sandra Forić – Maja Midžić: Gralis Speech-Korpus, Zagreb, Dezember 2007) oder **(2)** ein projekt- und korpusnahes Thema präsentiert wurde, wie etwa **(a)** auf der 34. Österreichische Linguistiktagung (Arno Wonisch: Das Forschungsprojekt „Die Unterschiede zwischen dem Bosnischen/ Bosniakischen, Kroatischen und Serbischen“, Klagenfurt, Dezember 2006), **(b)** auf dem I. Kongress der Wissenschafler Bosnien und Herzegowinas aus der Diaspora (Branko Tošović: Forschungsprojekt „Die Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen“, Sarajevo, September 2006), **(c)** bei der 36. Internationalen Slawistischen Tagung „Vukovi dani“ (Branko Tošović: Die grammatikalischen Unterschiede zwischen dem Serbischen, Kroatischen und Bosniakischen /Präliminarium/, Belgrad, September 2006), **(d)**

auf der 8. Internationalen wissenschaftlichen Konferenz „Zeit und Sprache“ (Branko Tošović: Die funktional-stilistischen Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen, Opole /Polen/, September 2006) und **(e)** im Rahmen eines Gastvortrages am Institut für slawische Philologie der Universität Śląsk (Branko Tošović: Die Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen, Katowice /Polen/, Dezember 2006). Im Zuge dieses Aufenthaltes in Katowice wurde mit der polnischen Seite vereinbart, ein spezielles Korpus für die Aktionsarten in den slawischen Sprachen zu entwickeln, das in seinem Anfangsstadium die Sprachen BKS, polnisch und russisch umfassen soll.

Für die Erstellung des BKS-Korpus wurde aus einem Teil der vom FWF für das Projekt „Die Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen“) genehmigten finanziellen Mittel die erforderliche technische Ausstattung angeschafft (vier PCs, ein Laserdrucker, zwei Scanner, eine Leinwand, vier Diktiergeräte, ein LCD-Fernseher u. a.), und von der Firma Linguatec aus München erging als Geschenk ein Laptop. Seitens des Institutes für Slawistik wurde der Raum 1.228 zur Verfügung gestellt, in dem die angeführte technische Ausrüstung untergebracht wurde und der zur Weiterentwicklung des Gralis-Korpus und zur Durchführung des genannten Projektes dient.

3. Das Gralis-Korporarium stellt ein System mehrerer Subkorpora dar, die schriftliche und mündliche (Video- und Audio-)Aufnahmen umfassen, wobei eine Unterteilung in das Text- und das Speech-Korpus erfolgt.

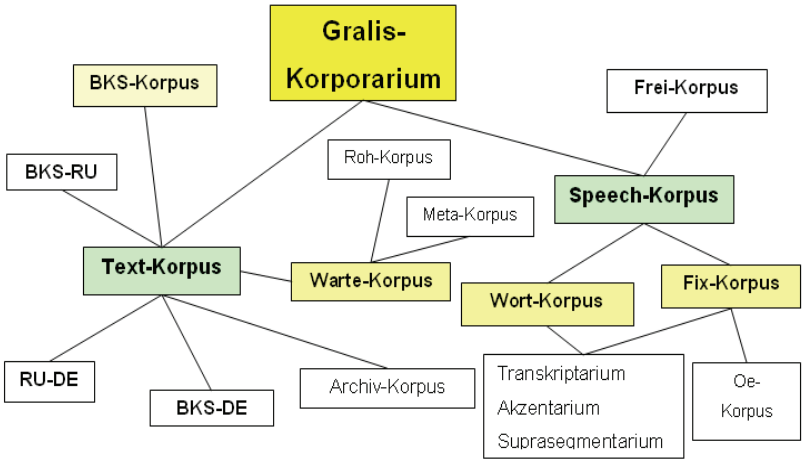


Abb. 4: Die Struktur des Gralis-Korporariums

4. Beim Text-Korpus handelt es sich um eine Online-Sammlung paralleler Texte für verschiedene slawische Sprachen. Fertig gestellt konnte bislang das Korpus für die Sprachen bosnisch/bosniakisch, kroatisch und serbisch werden, wobei dieses Korpus rund zwei Millionen Tokens beinhaltet. Gegenwärtig wird an der Erstellung eines solchen Korpus für weiter slawische Sprachen gearbeitet. Das Ziel des Gralis-Korpus liegt darin, ein Korpus zu erstellen, das **(a)** von keinerlei äußeren Faktoren abhängig ist, **(b)** in der Lage sein wird, mit der Geschwindigkeit und der Qualität der Informationstechnologien Schritt zu halten und **(c)** laufend weiterentwickelt, vervollständigt und verbessert werden kann.

Im Unterschied zur durchaus großen Zahl an einsprachigen Korpora trifft man sowohl innerhalb der Slawia als auch in allen anderen Philologien auf eine wesentlich kleinere Zahl an Parallelkorpora für zwei oder gar mehrere Sprachen. Dieses Ungleichgewicht liegt neben dem primären Interesse der Korpuslinguistik an der eigenen Sprache vor allem auch im technisch unvergleichlich anspruchsvolleren Prozess der Entwicklung von Parallelkorpora begründet. Doch gerade im Interesse einer ausgewogenen und komplexen Untersuchung der Übereinstimmungen, Ähnlichkeiten und Unterschiede zwischen nahe verwandten Sprachen (wie eben im Falle von B, K, S) erschien es unabdingbar, ein Korpus mit mehreren Sprachen zu entwickeln.

5. Nach Abschluss aller Arbeitsschritte wird das Gralis Text-Korpus aus dem Archiv-Korpus und dem Warte-Korpus bestehen. Das Archiv-Korpus beinhaltet Originaltexte, so wie sie von HerausgeberInnen, Redaktionen, ProduzentInnen, FilmvertreiberInnen, AutorInnen, ÜbersetzerInnen und RechtsnachfolgerInnen verstorbener TrägerInnen von Autorenrechten erhalten werden (ist einzig dem Leiter und dem Koordinator des Korpus zugänglich), wobei eine Einsichtnahme in das Material dieses Subkorpus nicht möglich ist. Die Texte im diesen Korpus verfügen über folgende Metainformationen: Quelle des Originals (Verlag, Zeitschriftenredaktion, Autor, ÜbersetzerIn, Link), Kurztitel, Sammeltitel (z. B. Zeitungen eines Monats), Datum und Ort der Herausgabe, Datum des Einfügens in das Archiv-Korpus, Art des Originals (gemäß ISO 639-2, ISO TO 37/SC2), Identifikationsnummer, Original oder Übersetzung (Name des Übersetzers/der Übersetzerin), ISBN-Nummer und ISSN-Nummer (fakultativ), Formatierung (Übereinstimmung der Absätze, Grafik, diakritische Zeichen) sowie willkürlicher Kommentar.

Das Warte-Korpus umfasst Originaltexte, die aus dem Internet zur weiteren Bearbeitung ausgewählt werden (http://www-gewi.kfunigraz.ac.at/gralis/0.Projektarium/BKS-Forum/BKS-Forum_Index.htm) und die einzig den am Korpus mitarbeitenden Personen zugänglich sind. Für die Erstellung des Warte-Korpus wird um keine Urheberrechte angesucht.

Die Arbeit an sämtlichen Subkorpora erfolgt parallel in verläuft in zwei Phasen: In der ersten werden Texte gesammelt und grob bearbeitet, um sie in das

nichtlemmatisierte Warte-Korpus einzustellen. In der zweiten Phase wird das lemmatisierte Korpus erstellt, indem repräsentative Textstellen aus dem Warte-Koprus elektronisch bearbeitet und in das Korpus eingefügt werden.

Eine weitere Untergliederung des Warte-Koprus führt zu zwei Subkorpora, die als Roh- und Meta-Korpus bezeichnet werden. Ersterer umfasst Texte aus dem Internet, die in zumindest zwei sprachlichen Versionen vorliegen, während zweiter eine Sammlung von Texten und Artikeln zur globalen Thematik des Projektes beinhaltet (bis dato liegt das Meta-Korpus einzig zum Thema „Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen“ vor).

Im Text-Korpus werden drei Sorten von Texten proportional und ausgeglichen inkludiert: 1. Originaltexte, 2. modifizierte (adaptierte) Texte und 3. übersetzte Texte. Das Gralis-Korpus wird aus einem schriftlichen und einem mündlichen Subkorpus bestehen, deren Verhältnis sich auf 90%:10% beläuft. Der Umfang von Texten hängt von dessen funktionalstilistischer und genremäßiger Zugehörigkeit ab. Um eine Ausgewogenheit zu erreichen, werden manche Texte (z. B. Romane) nur in Auszügen herangezogen.

Abhängig von der Lösung der Urheberrechtsfrage kann das Gralis Text-Korpus (a) eine begrenzte Zeit (z. B. ein Jahr) zugänglich sein, worüber man ein Vertrag mit den InteressentInnen schließen würde und (b) von einer begrenzten Anzahl von Personen genutzt werden (wie etwa MitarbeiterInnen des Instituts der Slawistik, inskribierten Studierenden, DiplomandInnen und DoktorandInnen, Studierenden, die den Unterricht aus Fachgebieten besuchen, der in Verbindung mit dem Thema Korpus oder Korpuslinguistik steht, Gästen des Instituts, Angehörigen anderer Institute und Fakultäten usw.).

Das Gralis Text-Korpus verfügt über drei Arten der Annotation: 1. eine metatextuelle, 2. eine extralinguistische und 3. eine linguistische (morphologische, orthoepische, semantische, stilistische und syntaktische), wobei die metatextuelle Annotation Informationen zu Titel, Kapitel und Absatz bietet.

Die extralinguistische Annotation verfügt über folgende Komponenten – **(1)** AutorIn: individuelle(r) AutorIn (Vor- und Nachname), kollektive(r) AutorIn (Vor- und Nachname), fingierte(r) AutorIn (Vor- und Nachname), Pseudonym, unbekannte(r) AutorIn (NN), Geburtsdatum (oder ungefähres Alter), Geschlecht, Nationalität, Konfession, Herkunft (Staat, Land, Stadt), Berufsfeld (Kunst, Publizistik, Wissenschaft, Recht usw.); **(2)** Editionsangaben: Umfang des Textes (Seitenzahl), Zeit des Entstehens des Textes, Ort des Entstehens des Textes, HerausgeberIn; Angaben zur Sprache, zur regionalen Variante, Schrift, Übersetzung (ÜbersetzerIn); **(3)** textuelle Angaben: Medium (schriftlich, mündlich), Textdomäne (Recht, Psychologie usw.), funktionaler Stil (literarisch-künstlerisch, publizistisch, wissenschaftlich, administrativ, umgangssprachlich), „Unterstil“ (informativ, analytisch, populärwissenschaftlich), Genre (Prosa, Poesie,

Drama, Artikel, Dissertation), Herkunft des Textes (Buch, Radiosendung, Zeitungsbeilage usw.), Typ der Sprachkommunikation (Monolog, Dialog, Gespräch, Vortrag); **(4)** inhaltliche Angaben: Thema (z. B. Kampf gegen Drogenmissbrauch, Kochrezept usw.), Chronotop (welche Zeit und welcher Ort werden im Text behandelt); **(5)** strukturelle Angaben: Art der Formatierung, Reim (falls vorhanden) und **(6)** kommunikatorische Angaben (für wen wurde der Text verfasst): für welche Altersgruppe, für Personen welchen Bildungsniveaus.

Die linguistische Annotation umfasst die Hervorhebung von Sätzen, Syntagmen und Wörtern, wobei zwischen folgenden weiterführenden Annotationsschritten unterschieden wird: **(a)** morphologische Annotation: nach morphosyntaktischen Kategorien; **(b)** orthoepische Annotation: nach der Art des Akzents (lang steigend, lang fallend, kurz steigend, kurz fallend, Länge); **(c)** semantische Annotation: gemäß dem Programm WortNet; **(d)** stilistische Annotation: nach der Art des Stils, der Art des funktionalen Stils (literarisch-künstlerisch, publizistisch, wissenschaftlich, administrativ, umgangssprachlich) und **(e)** syntaktische Annotation gemäß dem syntaktischen Baum der Abhängigkeiten.

Diese Annotationsschritte werden in mehreren Phasen erfolgen, wobei zuerst die metatextuelle Annotation, in einer zweiten Phase die morphologische und orthoepische, in einer dritten die semantische und stilistische sowie schließlich in einer vierten Phase die syntaktische Annotation durchgeführt werden. Morphosyntaktische Homographie soll händisch entfernt werden.

6. Bei der Textverarbeitung werden zwei grundlegende Verfahren zur Anwendung gebracht, nämlich die Segmentierung und das Alignieren. Im Zuge des Segmentierungsschrittes wird jeder Text in Absätze und Sätze unterteilt, woraufhin die Segmente angeglichen werden. Auf diese Weise wird eine strukturelle Übereinstimmung zwischen den Texten der untersuchten Sprachen hergestellt, sodass ein angeglichenes Parallelkorpus entsteht. Durch diese Arbeitsschritte werden die Wechselbeziehungen zwischen zwei oder mehreren sprachlichen Textversionen mit dem gleichen Inhalt dargestellt, woraufhin eine linguistische Analyse erfolgen und ein (alphabetisches) Frequenzwörterbuch ausgearbeitet werden kann.

Das Problem bei der Segmentierung und Alignierung von Texten liegt darin, dass beide Arbeitsschritte doppelt (sofern es sich um einen Text in zwei Sprachen handelt) oder sogar dreifach (wenn ein Text in drei Versionen in Frage kommt) durchgeführt werden müssen. In der Anfangsphase der Angleichung wird folgendes Modell zwischensprachlicher Beziehungen überprüft, angewandt oder modifiziert (**A** – **B** – **C**): (1) ein Satz der Sprache **A** hat als Äquivalent einen Satz mit übereinstimmenden Grenzen in den Sprachen **B**, **C** (Beziehung 1:1:1); (2) ein Satz der Sprache **A** hat als Äquivalent einen Satz mit nichtübereinstimmenden Grenzen in den Sprachen **B**, **C** (Beziehung 1:1:1); (3) ein Satz der Sprache **A** hat

als Äquivalent zwei (oder mehr) Sätze in den Sprachen **B**, **C** (Beziehung 1:1:2, 1:2:1 oder 2:1:1); (4) ein Satz der Sprache **A** hat keinen Äquivalent in den Sprachen **B**, **C** (Beziehung 1:1:0, 1:0:1 oder 0:1:1).

Texte, die direkte Übersetzungen darstellen, werden nach folgenden Kombinationen angeglichen: Dem Original entspricht eine authentische Übersetzung (amtliche Dokumente mit gleichwertiger Rechtskraft); dem Original entspricht eine Übersetzung des Autors/der Autorin bzw. eine autorisierte Übersetzung (eine beauftragte Übersetzung); dem Original entspricht eine maschinelle Übersetzung; dem Original entspricht keine Übersetzung, sondern ein modifizierter Text.

Das Gralis-Korpus soll in höchstmöglichem Maße dem Anspruch der Repräsentativität (zur Filterung zuverlässiger Informationen) und der Ausgewogenheit (zu einer adäquaten Darstellung der Differenzierung vor allem in funktionalstilistischer Hinsicht) gerecht werden. Als theoretische Grundlage für die typologische Einteilung der Texte dient dabei das Buch „Die funktionalen Stile“ (Tošović 2002). Gemäß dieser Konzeption wird das Gralis-Korpus in die fünf funktionalen Stile (literarisch-künstlerisch, publizistisch, wissenschaftlich, administrativ und umgangssprachlich) unterteilt.

Die Weiterentwicklung des Gralis-Korpus geht wie folgt vor sich: 1. quantitative Ergänzung durch neue Texte und Inhalte, 2. qualitative Verbesserung (tiefere und umfangreichere Annotation), 3. formale Verbesserungen (Erneuerung des Web-Designs), 4. funktionale Beschleunigung (besseres Such- und Findsystem) und 5. Weiterentwicklung der Programme (Anwendung neuer Softwarepakete).

Angesichts dessen, dass die Qualität jedes Korpus durch (a) die Tiefe und den Umfang der Annotation, (b) die Such- und Auffindmöglichkeiten, (c) die Repräsentativität, Proportionalität und Ausgewogenheit sowie (d) die Zugänglichkeit bestimmt wird, wird diesen Faktoren bei der Ausarbeitung und stetigen Weiterentwicklung des Korpus umfassend Rechnung getragen werden.[5]

Für eine Übertragung der Urheberrechte wird um diese bei Verlagen, Zeitungs- und Zeitschriftenredaktionen, FilmproduzentInnen und Verleihen, AutorInnen gedruckter und elektronischer Versionen von Texten, ÜbersetzerInnen oder – sofern sie nicht mehr am Leben sind – rechtmäßigen ErbInnen angesucht.

7. Ein Teil des Gralis Text-Korpus stellt das BKS-Korpus dar, bei dem es sich um ein paralleles informationell-wissenschaftliches System für das Bosnische/ Bosniakische, Kroatische und Serbische handelt, das aus zumindest in zwei

[5] Zur Nutzung des Gralis Text-Korpus siehe den Beitrag von Arno Wonisch in diesem Band.

Versionen vorliegenden Texten besteht (B und K, B und S, K und S). Das Ziel des BKS-Korpus liegt darin, in einer möglichst tiefen und umfassenden Untersuchung der Übereinstimmungen, Ähnlichkeiten und Unterschiede zwischen dem Bosnischen/Bosniakischen, 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. Basierend auf diesem Korpus könnte man mit der Ausarbeitung eines Programms für eine automatische Bestimmung des Grades der Nähe zwischen diesen Sprachen bzw. für eine Messung der typologischen Distanz beginnen. Weiters soll mithilfe des Korpus umfassendes Material für das Verfassen 1) eines komplexen korrelativen Wörterbuches der Sprachen B, K, S in einer gedruckten und einer Online-Version, 2) korrelativer Grammatiken des B, K, S und schließlich 3) eines Lehrbuchs des B, K, S zusammengetragen, aufbereitet und ausgewertet werden.

Das Gralis BKS-Korpus wendet sich an Fachleute für das BKS und LinguistInnen allgemeinen Profils (vor allem auf dem Gebiet der allgemeinen, der Systemlinguistik und der Soziolinguistik) sowie an all jene, die an den intralinguistischen, interlinguistischen und extralinguistischen Beziehungen zwischen dem B, K, S Interesse bekunden. Es kann breit und zweckmäßig im Unterricht und dabei vor allem an Hochschulen zum Einsatz gebracht werden, wobei es auch all jenen von Nutzen sein wird, die in der Praxis mit den Problemen des B, K, S konfrontiert sind (LektorInnen, Filmschaffenden, PolitikerInnen u. a.). Das Korpus stellt in erster Linie ein Parallelkorpus des Standardbosnischen, des Standardkroatischen und des Standardserbischen dar. Aus diesem Grund werden in einer ersten Phase nach dem Jahr 1991 verfasste Texte ausgewählt und bearbeitet. In einer zweiten Phase wird mit Texten gearbeitet, die zwischen 1981 und 1990 entstanden sind, in einer dritten Phase folgen Texte aus den Jahren 1961 bis 1980 und in einer vierten Phase Texte, die zwischen 1941 und 1960 erstellt wurden.

Die Entwicklung des Gralis-Korpus erfolgt gemäß den gängigsten Standards (z. B. TEI), um dadurch eine Kompatibilität und eine Vergleichbarkeit mit ähnlichen Korpora sowie breite Anwendungsmöglichkeiten zu erzielen. Die Arbeit am Gralis-Korpus ist einerseits eine einmalige (durch die Erstellung einer Online-Version) und andererseits eine laufend durchzuführende (ständige Ergänzungen, Verbesserungen und Vertiefungen).

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. In naher Zukunft soll die Verwaltung der Textdaten im Gralis Text-Korpus, die derzeit noch Filesystem-basiert erfolgt, auf ein sogenanntes Asset Management-

System (AMS) umgestellt werden. Korpustexte, aber auch zugehörige Audio-, Video- und beschreibende Metadaten, wie sie in einem multimodalen Korpus in einer Vielzahl vorhanden sind, können mittels eines solchen Frameworks einfach verwaltet und in webbasierten Workflows bearbeitet werden. Interessierte LeserInnen seien auf den Beitrag von Hubert Stigler in diesem Band verwiesen, der die Möglichkeiten dieser Umgebung detailliert darstellt.

8. Einen weiteren Teil des Gralis-Korpus stellt das Speech-Korpus dar. Es handelt sich dabei um eine Online-Sammlung von Audiomaterial (gegenwärtig vorerst nur für das Bosnische/Bosniakische, Kroatische und Serbische), die aus drei Subkorpora – dem Wort-, Fix- und Frei-Korpus – besteht.

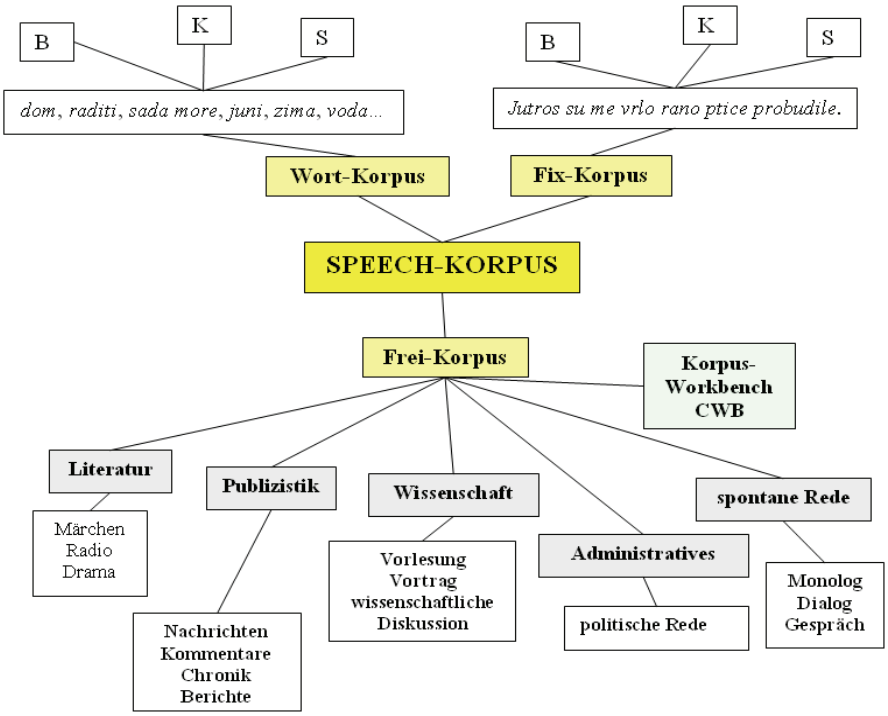


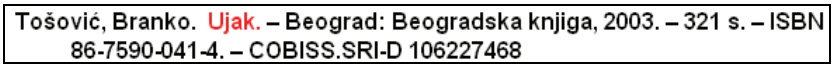
Abb. 5: Die Struktur des Gralis Speech-Korpus

Es sei an dieser Stelle vorab darauf hingewiesen, dass das Wort-Korpus aus Aufnahmen verlesener Wortlisten besteht und es sich beim Fix-Korpus um Aufnahmen kürzerer Texte (der häufig verlesene Text „Jutro“ umfasst 18 Sätze) handelt. Genauere Erklärungen zu diesen Subkorpora (Wort- und Fix-Korpus im Rahmen des Gralis Speech-Korpus) finden sich in weiteren Beiträgen in diesem Kapitel.

Im Rahmen des Speech-Korpus wird auch ein Phonokorpus für die deutsche Sprache in Österreich erstellt (Oe-Korpus), das dazu dienen soll, mittels einer typologischen Untersuchung die Aussprache in Deutschland und Österreich zu vergleichen und die Übereinstimmungen, Ähnlichkeiten und Unterschiede zwischen nahen Sprachen und ihren Varietäten zu erheben. Das Oe-Koprus wird gemäß einer Vereinbarung zwischen der Firma „Linguatec Sprachtechnologien GmbH“ aus München und dem Leiter des Forschungsprojektes „Die Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen“ entwickelt, wobei den Gegenstand der Zusammenarbeit Aufnahmen österreichischer Sprechender im Sinne einer Erhöhung der Qualität von Spracherkennung für das Deutsche darstellen. Den Output der Aufnahmen bilden Audiodateien im wav-Format mit jeweils 200 Sätzen aus insgesamt 24 unterschiedlichen Skripts, wobei zu jeder aufgenommenen Person wesentliche Metadaten erfasst werden. Die Sprachaufnahmen werden mit dem von Linguatec entwickelten Software-Tool „npcmrec“ vorgenommen und wurden mit Ende Jännner 2008 abgeschlossen.

Das dritte Subkorpus im Rahmen des Gralis Speech-Korpus bildet schließlich das Frei-Korpus, 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 (Tošović 2003) abgedruckt wurde. Eine Suche im Frei-Korpus erfolgt analog zu jener im Text-Korpus, wobei sich die Findstellen wie folgt darstellen:

Am oberen Ende des Suchfensters befindet sich der Verweis auf die Quelle in Form eines Kurztitels (Ujak), auf den ein Pfeil folgt. Klickt man auf den Satz, erhält man die Information zur gesamten bibliographischen Quelle:



Mit einem Klick auf den Satz erhält man weiters auch die Möglichkeit, diesen zu hören. Jeder segmentierte Satz ist mit Audiodateien in zwei Formaten – wav und mp3 – versehen. Die Aufnahme im wav-Format dient für die akustische Analyse

und ist (auf Grund des großen Datenumfanges) online nicht zugänglich, sodass in Gralis ausschließlich Aufnahmen im mp3-Format eingestellt werden.

Einen wesentlichen Teil des Frei-Korpus bilden Radio- und TV-Aufnahmen, deren Besonderheit darin liegt, dass sie textuelle, akustische und visuelle Informationen beinhalten. Im Rahmen der Aktivitäten zur Entwicklung des Frei-Korpus wurden z. B. am selben Tag und zur selben Zeit (19.30–20.00 Uhr) die TV-Nachrichten des serbischen, kroatischen und bosnisch-herzegowinischen Fernsehens aufgenommen, die in einem ersten Arbeitsschritt transkribiert wurden. Die gesamte Information (Ton, Bild und Text) wurde sodann in Sätze segmentiert und auf den Server überspielt. Das Ziel lag dabei darin, eine Synchronisation zwischen Text Ton und Bild herzustellen.

9. Im Rahmen des Gralis-Komplementariums kam es zur Ausarbeitung mehrerer Datenbanken, die entweder direkt aus den Subkorpora entstanden oder für ein Funktionieren des Gralis-Korpus dienen. Das Gralis-Komplementarium stellt ein Programmsystem zur Sammlung und Bearbeitung von Material für sämtliche Subkorpora dar.

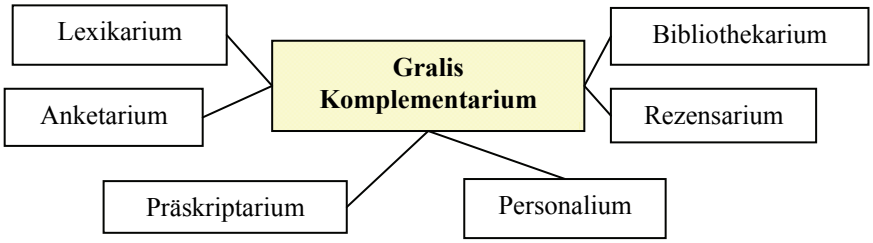


Abb. 6: Die Struktur des Gralis-Komplementariums

Einen weiteren Teil des Gralis-Korpus stellen das Gralis-Lexikarium, das Gralis-Anketarium, das Gralis-Bibliothekarium, das Gralis-Präskriptarium, das Gralis-Personalium und das Gralis-Rezensarium dar.

Beim Gralis-Lexikarium handelt es sich um ein Online-Wörterbuch, das auf den anderen Teilen des Gralis-Komplementarium fußt und für Forschungen zur lexikalischen Struktur slawischer Sprachen dient. Das Gralis-Lexikarium stellt ein internationales Forschungsprojekt zur Ausarbeitung eines bidirektionalen Online-Wörterbuches für die Sprachen deutsch ↔ bosnisch/bosniakisch, kroatisch, montenegrinisch und serbisch mit späterer Ausweitung auf andere slawische Sprachen dar und befand sich zur Zeit der Drucklegung dieses Bandes in der Entwicklungsphase.

Eine weitere Komponente des Gralis-Komplementariums, das Gralis-Anketarium, dient zur Sammlung von Quellen mittels Versendung von Online-

Umfragen, wobei diese in jeder beliebigen Sprache erstellt werden können. Das Anketarium besteht aus drei Kategorien von Umfragen, von denen eine für wissenschaftliche Zwecke genutzt wird (Wissenschaftliche Umfragen), eine weitere Zwecken des Unterrichtes dient (Edukative Umfragen) und die dritte schließlich Umfragen zu unterschiedlichen Themenfeldern umfasst (Andere Umfragen). Als Benutzersprachen stehen die drei Studienrichtungssprachen des Institutes für Slawistik der Karl-Franzens-Universität Graz (BKS, russisch, slowenisch) und deutsch zur Verfügung. Genaueres zum Gralis-Anketarium siehe im Beitrag von Robert Thomann in diesem Kapitel.

Mithilfe des Gralis-Bibliothekariums erfolgt die Sammlung, Bearbeitung und Darstellung bibliographischer Angaben, die für alle mit dem Gralis-Korpus verbundenen Forschungsprojekte wie auch für edukative Zwecke unerlässlich sind. Ein Teil des Bibliothekariums ist für Sprachen mit lateinischer Schrift vorgesehen (Lat-Bibliothekarium), der andere Teil für all jene Sprachen, die sich des kyrillischen Alphabetes bedienen (Cyr-Bibliotehkaraium). Zum Gralis-Bibliothekarium siehe den Beitrag von Branko Tošović in diesem Kapitel.

Das Gralis-Präskriptarium dient zum Studium der Rechtschreibung slawischer Sprachen, indem es in sich die angebotenen standardologischen Lösungen mehrerer normativer Regelwerke für unterschiedliche Sprachen vereint. Mehr zum Gralis-Präskriptarium siehe im gleichnamigen Beitrag von Branko Tošović in diesem Kapitel.

Das Gralis-Personalium bietet eine Sammlung umfassender biographischer und bibliographischer Informationen zu Personen, die an den im Rahmen des Gralis-Portals beschriebenen Projekten mitarbeiten. Eine genauere Vorstellung dieses Programms erfolgt im Beitrag von Arno Wonisch.

Im Frühjahr des Jahres 2007 wurde das Gralis-Rezensarium in Betrieb genommen, mithilfe dessen eine Online-Beurteilung wissenschaftlicher Aufsätze vorgenommen werden kann, wobei die von den GutachterInnen getätigten Änderungsvorschläge automatisch an die VerfasserInnen und die Projektverantwortlichen übermittelt werden. Ein Teil des auf diese Weise entstehenden Sprachmaterials wird in das Text-Korpus integriert (Genre: Rezension; funktionaler Stil: wissenschaftlich). Genaueres zum Gralis-Rezensarium siehe im gleichnamigen Artikel von Stefan Kofler und Arno Wonisch.

10. Die Gralis-Tools setzen sich aus unterschiedlichen Programmen zusammen, die zur Bearbeitung textuellen und mündlichen Sprachmaterials und zu deren Aufnahme in das Gralis-Korpus dienen. Diese Tools umfassen (a) Programme zur Bearbeitung von Texten, (b) Programme zur Aufbereitung von Ton und Bild und (c) Programme zum Upload von Sprachmaterial auf Server.

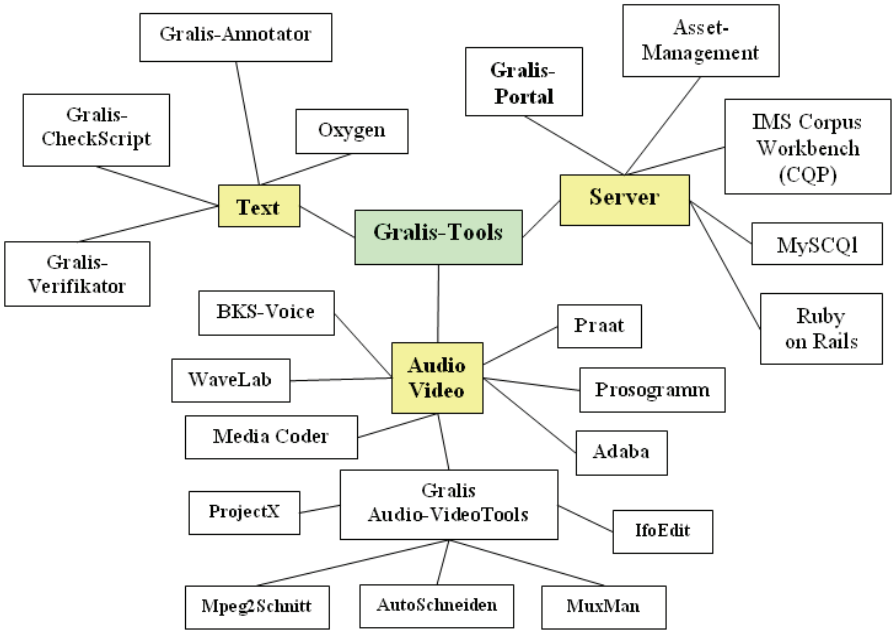


Abb. 7: Die Struktur der Gralis-Tools

Die Programme zur Bearbeitung von Texten bestehen aus dem Gralis-Annotator, dem Gralis-CheckSript, dem Gralis-Verifikator und dem Programm Oxygen. Es sei angemerkt, dass bei der Erstellung des Text-Korpus vor allem automatische Analysatoren (universelle, sprachunabhängige oder sprachspezifische) zur Anwendung kommen, wobei gegenwärtig ein von Hubert Stigler (Institut für Informationsverarbeitung in den Geisteswissenschaften – INIG) entwickeltes Programm zur Konvertierung ins XML-Format im Zentrum der Korpus-Arbeitsschritte steht. Dieses mit der Bezeichnung Gralis-Annotator versehene Programmpaket beruht auf den Prinzipien des Asset-Managements[6] besteht aus zwei Dateien, die die Namen gralis.dot und gralis.doc tragen, wobei gralis.doc eine Beschreibung der metatextuellen Annotationsmöglichkeiten („Druckformate“) und gralis.dot das eigentliche Programm darstellt. Das Programm basiert auf einem Word-Makro und definiert alle für eine metatextuelle Annotation erforderlichen Druckformate. Über einen Menüpunkt können in den Text Satzendmarker eingebracht werden, die gegebenenfalls manuell korrigiert (verschoben oder gelöscht) werden können. Zur einer möglichst korrekten Setzung dieser Satzendmarker wurden einige Heuristiken implementiert, wie etwa ein Auftreten

[6] Siehe dazu den Beitrag von Hubert Stigler in diesem Band.

von zwei Zeichen gefolgt von einem Punkt, wobei die Zeichen der Bedingung „kein Vokal“ entsprechen müssen, z. B.: mr., dr. u. a.). Auf Basis der Satzendmarker, die mit dem Text gespeichert werden, erstellt das Marko sodann die etikettierte xml-Datei (xml = Extensible Markup Language) im TEI-Standard (= text encoding initiative). BenutzerInnen erhalten Rückmeldung über die Anzahl der Absätze und Sätze im Text, wodurch die Editierung der Paralleltexte und die Fehlersuche erleichtert werden. Im Falle eines Nichtübereinstimmens der Anzahl von Absätzen und Sätzen besteht die Möglichkeit, mit dem ebenfalls von Hubert Stigler entwickelten Gralis-CheckSkript eine Valorisierung der mit dem Gralis-Annotator durchgeführten Arbeitsschritte vorzunehmen, wobei angezeigt wird, in welchen Absätzen Unterschiede hinsichtlich der Anzahl der Segmente (d. h. Sätze) vorliegen, die sodann zu beheben sind.

Ein weiteres Programm zur Überprüfung der Anzahl an Segmenten innerhalb von Absätzen wurde mit der Bezeichnung Gralis-Verifikator versehen und ermöglicht eine tabellarische Gegenüberstellung von Texten in jeweils zwei sprachlichen Versionen. Dabei wird durch das Abrufen eines Skripts neben den beiden, nach Absätzen gegliederten Tabellen für die Sprachversionen eine dritte Spalte hinzugefügt, in der eventuelle Abweichungen der Segmentanzahl ausgewiesen werden.

Nach Abschluss sämtlicher Arbeitsschritte zur Harmonisierung und Angleichung von Texten in mehreren sprachlichen Versionen folgt vor dem finalen Serverupload eine Gegenüberstellung im XML-Quellcode-Editor Oxygen, für den im November 2006 eine Lizenz erworben wurde.

Der nun folgende Arbeitsschritt liegt in der Transformation der fertig bearbeiteten Textdokumente auf einen Server, wofür eine Vertikalisierung vorzunehmen ist. Ein diesbezügliches Programm wurde im Herbst 2006 von Miloš Utvić entwickelt und kam bei früheren Arbeitsversionen der Textadaption zur Anwendung. Andere Applikationen, die in der Anfangsphase des Gralis Text-Korpus zur Anwendung kamen, stellten ebenfalls im Rahmen des Korpus der modernen serbischen Sprache an der Mathematischen Fakultät der Universität Belgrad entwickelte Technologien dar, von denen das Programm xAlign (in der Version von Duško Vitas) und das Parallelisierungsprogramm (tmx) von Ranka Stanković erwähnt seien.

Am 20. Dezember 2006 erfolgte schließlich die Inbetriebnahme des neuen Gralis-Annotators, wobei von Hubert Stigler eine kurze Einschulung abgehalten wurde, an der neben den am Korpus mitarbeitenden Personen auch Kurt Tiefenbacher teilnahm (zeichnete für eine erste Struktur des Gralis Text-Korpus verantwortlich). Zu diesem Zeitpunkt war von Hubert Stigler nach Konsultationen mit Tomaž Erjavec bereits ein Gesamtpaket geschnürt worden, das nach der Durchführung der Vertikalisierung einen einfach zu handhabenden Upload der

Texte auf den Server ermöglicht. Dank dieses Programmpaketes ist es nun möglich, mit einem verhältnismäßig geringen Zeitaufwand Texte in mehreren sprachlichen Versionen in mehreren Arbeitsschritten serverfertig aufzubereiten und auf Basis der IMS Corpus Workbench des Institutes für maschinelle Sprachverarbeitung der Universität Stuttgart[7] Teil des Gralis Text-Korpus werden zu lassen.

Eine weitere Komponente der Gralis-Tools bilden Programme zur Bearbeitung von Aufnahmen gesprochener Sprache, die sich primär aus dem Spracherkennungsprogramm BKS-Voice, den Gralis Audio-VideoTools und den Programmen WaveLab, Praat, Prosogramm und Adaba zusammensetzen.

Für eine Erkennung gesprochener Sprache ist für das Bosnische/Bosniakische, Kroatische und Serbische die Entwicklung eines Spracherkennungsprogramms namens BKS-Voice vorgesehen, deren Ziel darin liegen würde, a) ein effizienteres, rationelleres und billigeres Sammeln mündlicher Quellen zu ermöglichen und b) eine möglichst objektive Bestimmung der Konkordanzen, Ähnlichkeiten und Unterschiede der drei Sprachen in phonetisch-phonologischer Hinsicht und in der gesprochenen Sprache zu erleichtern. Die gebräuchlichsten und effizientesten Spracherkennungssysteme basieren auf den mathematischen Modellen von Markov und Gauß und der Methode von Basisvektoren zur Modellierung der akustischen und linguistischen Besonderheiten einer Sprache. Die Ausarbeitung der Methode und des Algorithmus erfolgt dabei durch gesammeltes Sprechmaterial mit einem Umfang von mindestens 5000 Wörtern. Die Entwicklung eines solchen Programms für das BKS wird in mehreren Etappen vor sich gehen: **1.** Analyse der phonetischen Struktur der Sprachen und Wahl der elementaren Einheiten zur Spracherkennung (Phonem, Allophon u. Ä.); **2.** Anlegen einer aus repräsentativem Material bestehenden akustischen Datenbank zur Modellierung akustischer Charakteristiken; **3.** Segmentierung der akustischen Datenbank in elementare Erkennungseinheiten; **4.** Wahl eines effizienten akustischen Vektors; **5.** Ausarbeitung des statistischen Modells (Markov-Modell) auf Basis vorhandener linguistischer Angaben und der segmentierten akustischen Datenbank (Transformationsblock Stimme → akustisches Symbol); **6.** Erstellen von Regeln der im Rahmen des gewählten statistischen Modells vorzunehmenden, allmählichen Umformung der elementaren Erkennungseinheiten in einen grammatikalisch korrekten Text (Transformationsblock Symbol → Wort). Zum Zeitpunkt der Drucklegung dieses Bandes werden die ersten Schritte zur Entwicklung von BKS-Voice im Rahmen einer Diplomarbeit von Alexander Friedl am Institut für Signalverarbeitung und Sprachkommunikation der Technischen

[7] Die Lizenz für diese Korpussoftware wurde im April 2006 erworben.

Universität Graz unter der Betreuung von Stefan Petrik und Gernot Kubin durchgeführt.

Die Gralis Audio-VideoTools stellen ein Skript zur Bündelung mehrerer Programme dar, mit denen Audio- und Videomaterial bearbeitet werden kann und dessen Hauptkomponenten die Programme ProjectX, Mpeg2Schnitt, MuxMan, IfoEdit und AutoSchneiden bilden. Genaueres siehe dazu im Beitrag von Boris Tošović in diesem Kapitel.

Abschließend seien zusammengefasst einige Programme genannt, die im Zuge der Bearbeitung von (mehrheitlich) Audiodateien laufend angewandt werden und sich im Sinne einer raschen und effizienten Korpuserstellung als zweckmäßig und zielführend erwiesen haben. Es sind dies die (in den weiteren Beiträgen dieses Kapitels genauer beschriebenen) Programme (1) WaveLab der Firma Steinberg zur Bearbeitung von digitalem Tonmaterial, dessen Version 6.0 aus dem Jahr 2006 vom Institut für Slawistik erworben wurde; (2) das am Institute of Phonetic Sciences an der Universität Amsterdam entwickelte Open-Source-Programm Praat, das für detaillierte akustische Analysen im Format wav herangezogen wird; (3) das an den Universitäten Genf und Brüssel ausgearbeitete Prosogramm, welches auf Praat basierend akustische Analysen zu Tonhöhenverlauf, Satzintonation und (im Falle des BKS) Akzentstruktur ermöglicht; (4) die Open-Source-Datenbank-software MySQL, die sämtlichen Datenbankstrukturen im Rahmen des Gralis Speech-Korpus, des Gralis-Bibliothekariums, Gralis-Präskriptariums, des Gralis-Personaliums und des in der Entwicklungsphase stehenden Gralis-Lexikariums zu Grunde liegt; (5) das Web-Framework Ruby on Rails zum schnellen Erstellen von Internetinhalten, das beim Gralis-Rezensarium zum Einsatz kommt und schließlich (6) Adaba (Aussprachedatenbank des Österreichischen Deutsch), die von Rudolf Muhr für die Zwecke des Projektes „Die Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen“ freundlicherweise zur Verfügung gestellt wurde und bei der Erstellung des Wort-Korpus im Rahmen des Gralis Speech-Korpus zum Einsatz kommt. Die Fertigstellung von Adaba stellt den Schlusspunkt eines über sechsjährigen Forschungsprojektes unter der Leitung von Rudolf Muhr dar, das die Erstellung eines phonetischen Korpus des Österreichischen Deutsch (ÖDt.) und darauf aufbauend die Ausarbeitung eines empirisch begründeten Aussprachewörterbuchs zum Ziel hatte.

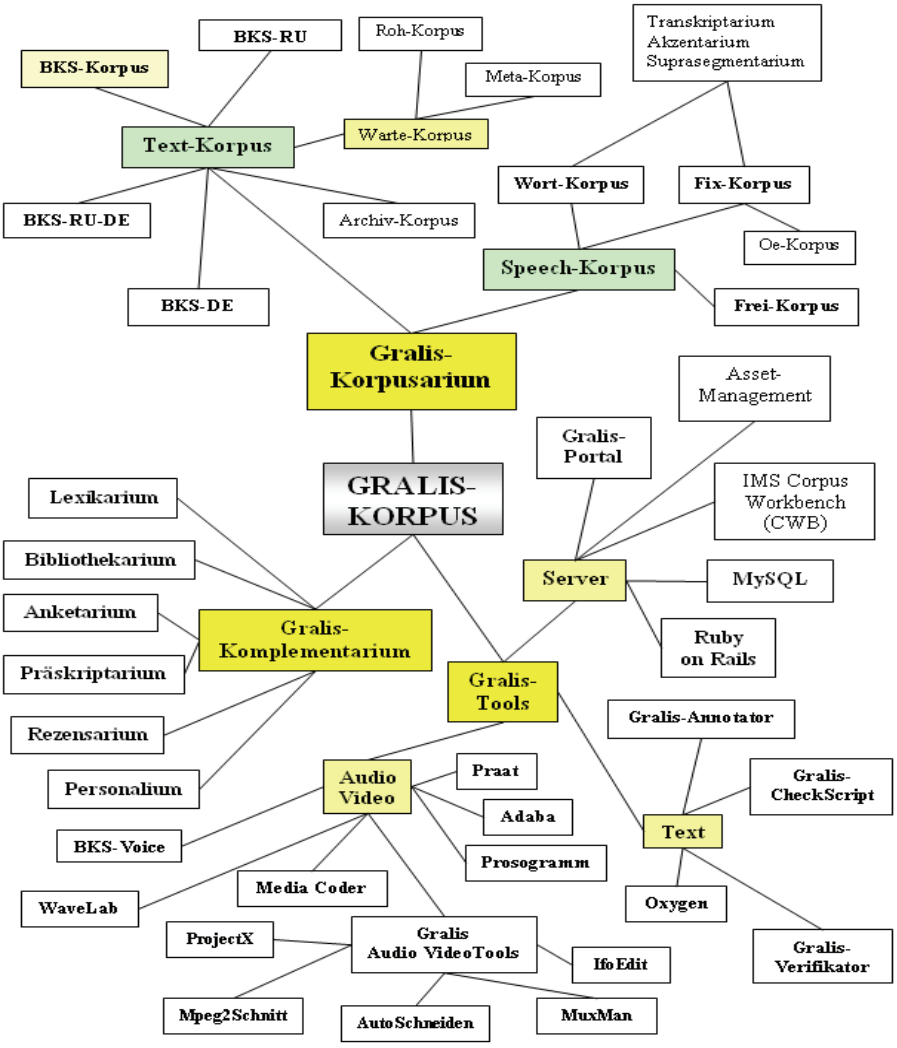


Abb. 8: Die Gesamtstruktur des Gralis-Korpus

Branko Tošović (Graz)

Gralis-Korpus

U prvom dijelu bloka tekstova posvećenom Gralis-Korpusu prezentirana je osnovna njegova koncepcija, način nastanka, pravci daljeg razvoja i osnovni dijelovi. U drugom dijelu predstavljen je Text-Korpus i Specchkorpus. U trećem dijelu govori se o tehničkom razvoju Gralis-Korpusa, snimanju, dekodiranju i preradi govornog materijala. Četvrti dio posvećen je programima za automatsko segmentiranje i analiziranje audio i video materijala (Gralis Audio-VideoTools), za prikupljanje istraživačke građe putem on-line anketiranja (Gralis-Anketarium) i za on-line recenziranje (Gralis-Rezensarium). Peti dio donosi priloge o programu izrađenom za prikupljanje i prezentiranje literature o slovenskim jezicima (Gralis-Bibliothekarium), programu za traženje i nalaženje podataka o učesnicima na projektima (Gralis-Personalium) i o programu za proučavanje međujezičkih pravopisnih korelacija (Gralis-Präskriptarium).

Gralis-Korpus je on-line informacijsko-analitički kompleks za prikupljanje, obradu i analizu tekstualne, govorne i vizuelne informacije izrađen radi sistemskog istraživanja slovenskih jezika. On predstavlja višejezičku, višedimenzionalnu i višenamjensku zbirku tekstova, audio i video snimaka kao i drugog prikupljenog i obrađenog materijala. Korpus je ime dobio ime po Gralisu – slavističkom portalu Univerziteta u Grazu (http://www-gewi.kfunigraz.ac.at/gralis).

Gralis-Korpus čine tri velika kompleksa – Gralis-Korporarium, Gralis-Komplementarium i Gralis-Tools. Gralis-Korporarium je satsavljen od više podkorpusa pisanih tekstova, govornih i video snimaka. On se sastoji od Text-Korpusa i Speech-Korpusa. Text-Korpus je on-line zbirka paralelnih tekstova za pojedine slovenske jezike. Za sada je gotov takav korpus za B, K, S i u datom trenutku sadrži oko dva miliona pojavnica. U toku je izrada sličnog korpusa za druge slovenske jezike. Drugi dio Gralis-Subkorpora je Speech-Korpus. On predstavlja on-line zbirku govornog materijala (u sadašnjoj fazi postoji samo za B, K, S). On je podijeljen na tri potkorpusa: Wort-Korpus, Fix-Korpus i Frei-Korpus. Wort-Korpus čine snimci dobijeni izgovaranjem izolovanih riječi. Fix-Korpus je zbirka audio materijala snimljenog na bazi čitanja manjih tekstova. Frei-Korpus je namijenjen za proučavanje spontanog govora. Pošto se za takav korpus ne postoje paralelni primjeri (svaka takva jezička realizacija predstavlja unikat za koji se ne može naći semantički identičan), već se može tražiti govorni iskaz koji odslikava istu situaciju (recimo, razgovor na pijaci, u restoranu) ili žanr (dijalog, pripovijedanje, diskusija, replika), ovaj potkorpus je izdvojen iz Speech-Korpusa zasnovanog na MySQL bazi podataka i uključen u Text-Korpus baziran na Workbench CWB. Značajan dio Frei-Korpusa činiće materijal dobijen iz radio i tv emisija. Njihova je specifičnost u tome da sadrže tekstualnu, slušnu i vizuelnu informaciju. U okviru Speech-Korpusa radi se na izradi govornog korpusa za njemački jezik u Austriji (Ö-Korpusa) radi tipološkog proučavanja podudarnosti, sličnosti i razlika između jezičkih varijeteta na njemačkom govornom području, što može biti značajno za tipološka proučavanja njemačko-slovenskih govornih korelacija.

Gralis-Komplementarium predstavlja sistem programa za prikupljanje i obradu materijala za sve podkorpuse, u prvom redu Text-Korpus i Speech-Koprus. U okviru Gralis-Korpusa razvijen je ili se nalazi u procesu izrade Gralis-Lexikarium, Gralis-Anketarium, Gralis-Bibliothekarium, Gralis-Präskriptarium, Gralis-Personalium i Gralis-Rezensarium. Gra-

lis-Lexikarium predstavlja sistem on-line rječnika, koji se naslanjaju na sve druge dijelove Gralis-Korpusa i služi za prezentiranje i proučavanje leksičke strukture slovenskih jezika. Gralis-Anketarium se koristi za dobijanje istraživačke građe putem anketiranja i tako je koncipiran da se ono može vršiti za bilo koji jezik. Gralis-Bibliothekarium služi za prikupljanje, obradu i prezentaciju bibliografskih podataka. Jedan njegov dio je namijenjen za jezike koji su služe latinicom (L-Bibliothekarium), drugi ćirilicom (C-Bibliotehkaraium). Gralis-Präskriptarium je namijenjen za proučavanje pravopisnih međujezičkih korelacija. U prvoj fazi radi se na izradi BKS-Präskriptariuma. Gralis-Personalium daje informaciju o učesnicima na projektima. Gralis-Resensarium služi za on-line recenziranje radova. Dio prikupljenog materijala koristiće se za Text-korpus (u žanru recenzije i naučnom stilu).

Gralis-Tools čine sistemi za preradu pisane i govorne građe radi njihova uključivanja u Gralis-Korpus. On se sastoji od (a) programa za obradu pisanog teksta, (b) programa za obradu glasa i slike te (c) server-programa. Programe za obradu pisanog teksta čini Gralis-Annotator, Gralis-CheckSript i Gralis-Verifikator. Gralis-Annotator služi za automatsko markiranje kraja rečenica radi segmentiranja teksta i povezivanja dijelova različitih jezičkih verzija po sistemu rečenica A_1 – rečenica A_2 (tzv. paralelizovanje). Gralis-CheckSript koristi se za valorizaciju procedura urađenih u okviru Gralis-Annotatora. Gralis-Verifikator služi za provjeru da li se paralelizovane rečenice nalaze u odnosu 1 : 1. Programi za obradu glasa objedinjuje Gralis-AudioVideoTools. On predstavlja skript koji povezuje nekoliko programa za obradu audio, video i SAT podataka, prije svega ProjectX, Mpeg2Schnitt, MuxMan, IfoEdit i AutoSchneiden. Radi prepoznavanja BKS-govora predviđena je izrada BKS-Voice**.** Kao server-programi koriste se IMS Corpus Workbench (CQP), MySql, Ruby on Rails i Asset-Management.

Branko Tošović
Institut für Slawistik
Karl-Franzens-Universität Graz
Merangasse 70
8010 Graz
Österreich
Tel.: +43/316/380 2522
Fax: +43/316/380 9773
branko.tosovic@uni-graz.at
http://www-gewi.uni-graz.at/gralis/