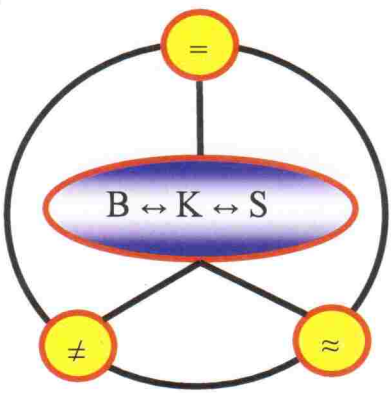
Branko Tošović (Hg.)

Die Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen 1/3

Das Gralis-Korpus



Sonderdruck, S. 724–827

Slawische Sprachkorrelationen 1

LIT

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 bibliographischen 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 korpus, 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: Pro-

[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.

jektarium, Korpusarium, Educarium, Gralisarium, Grazer Slawisticarium und Operarium.

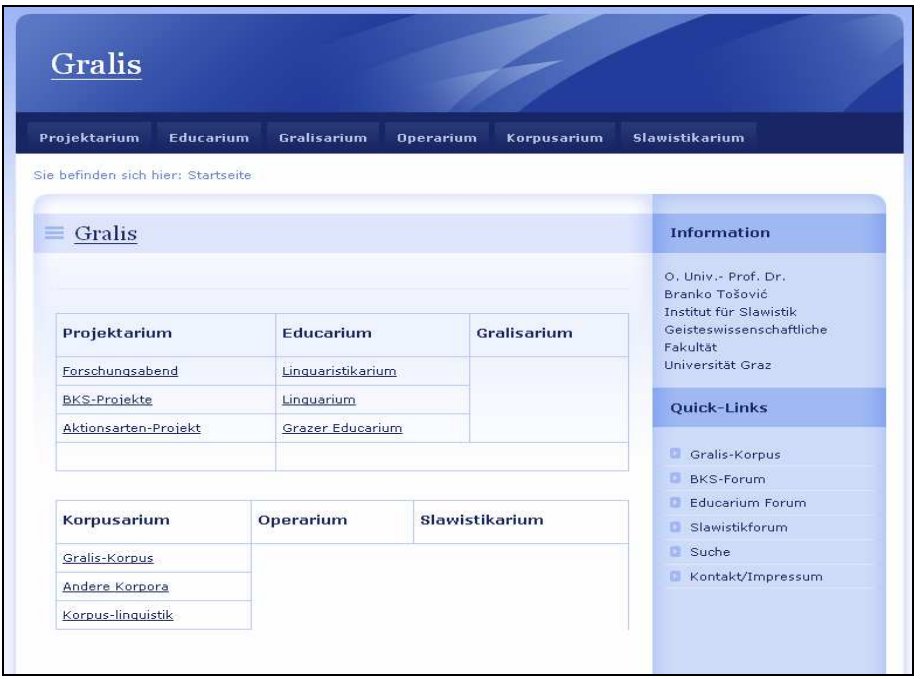


Abb. 1: Die Gralis-Startseite

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 Linguisticarium 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.

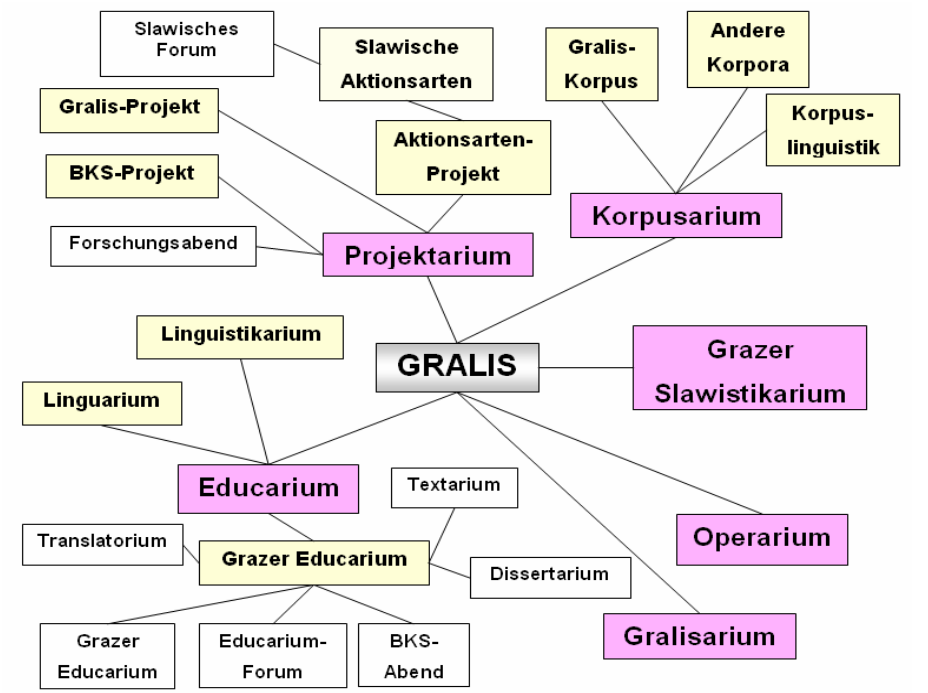


Abb. 2: Die Struktur von Gralis

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.

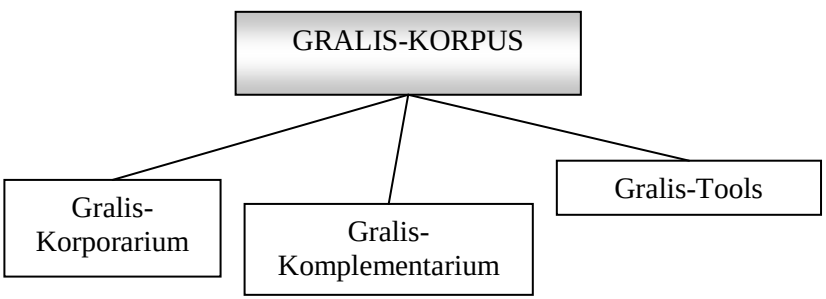


Abb. 3: Die Struktur des Gralis-Korpus

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 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

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

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 National-

korpus (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 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.

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

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.

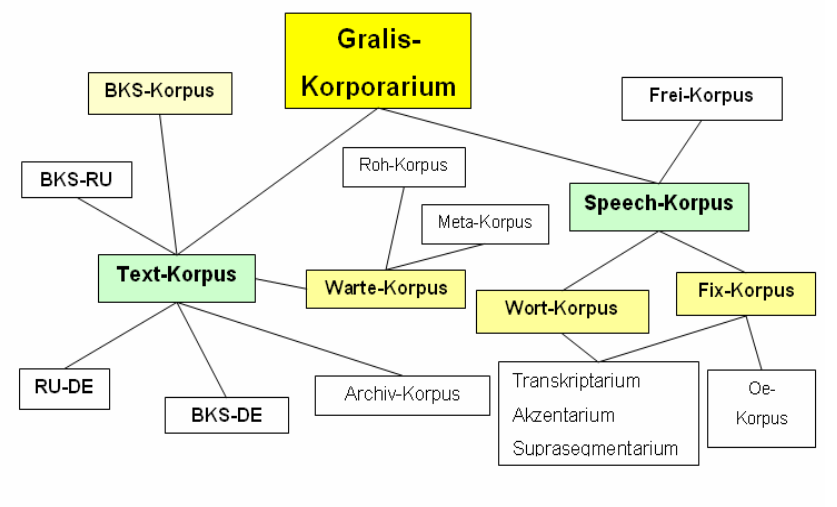


Abb. 4: Die Struktur des Gralis-Korporariums

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 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 Speku-

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

lationen 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.

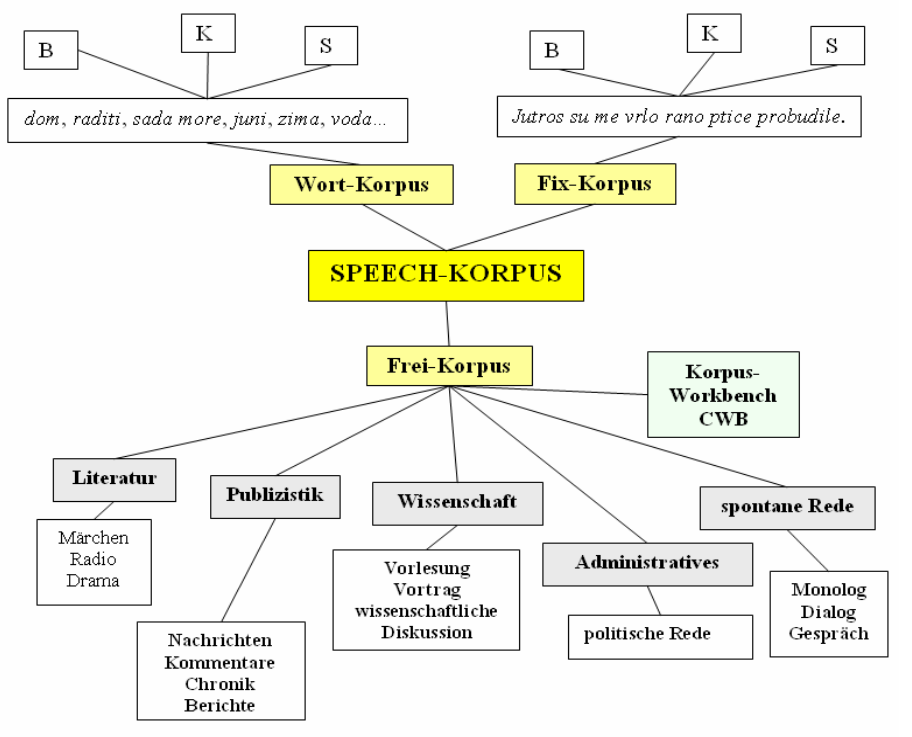


Abb. 5: Die Struktur des Gralis Speech-Korpus

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 Forschungspro-

jektes „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:

Tošović, Branko. Ujak. – Beograd: Beogradska knjiga, 2003. – 321 s. – ISBN 86-7590-041-4. – COBISS.SRI-D 106227468

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.

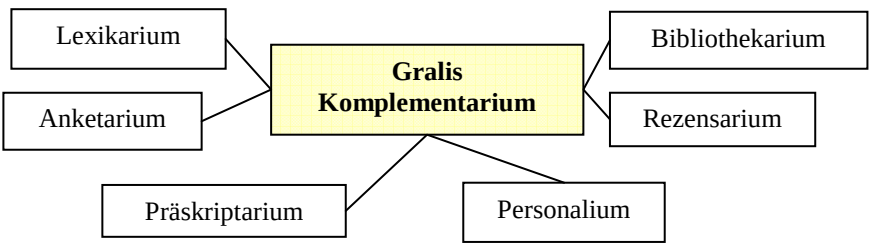


Abb. 6: Die Struktur des Gralis-Komplementariums

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.

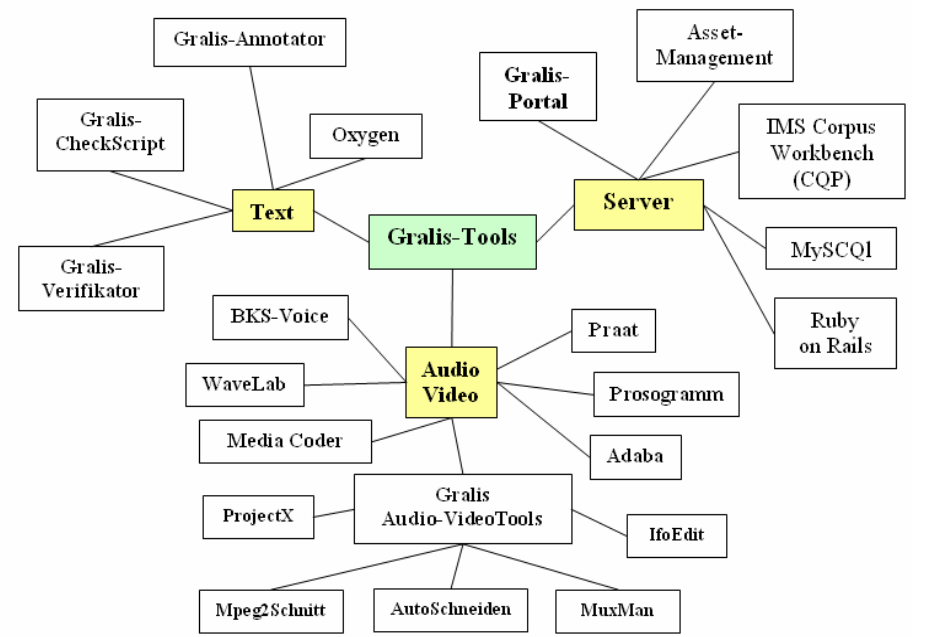


Abb. 7: Die Struktur der Gralis-Tools

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 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

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

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 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,

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

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-Datenbanksoftware 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.

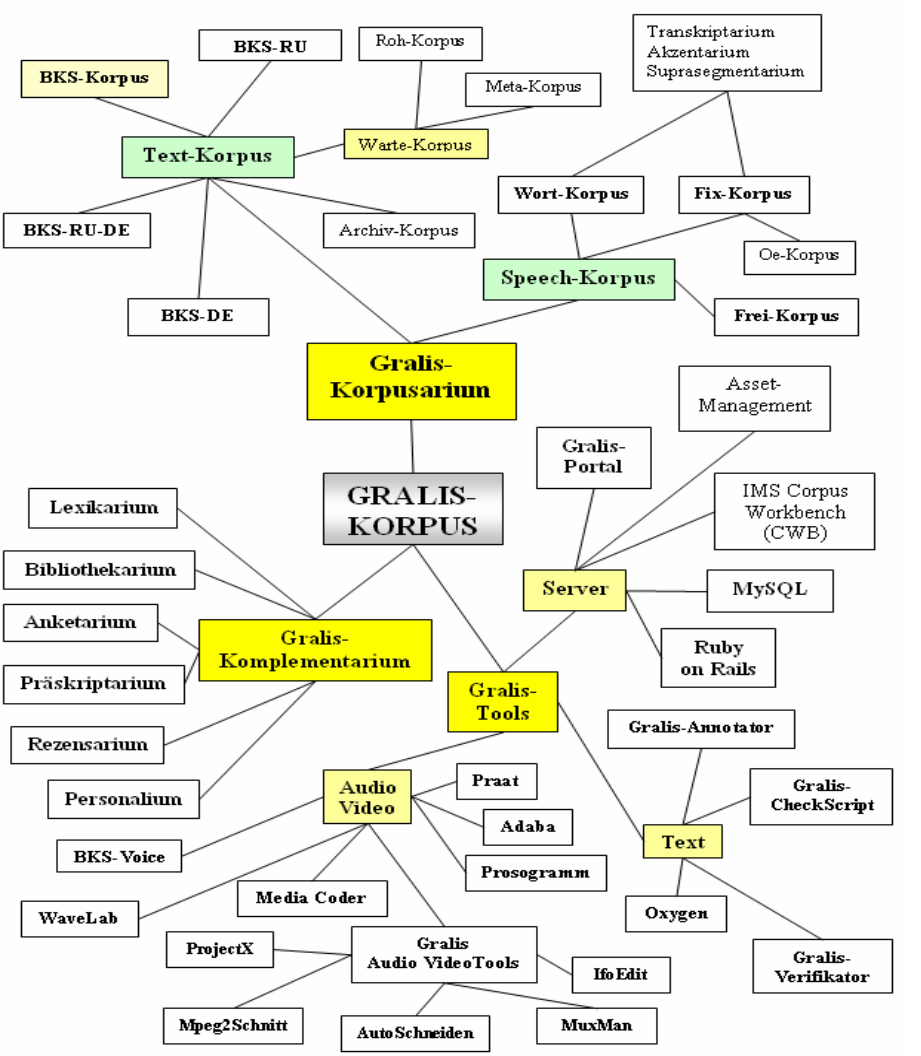


Abb. 8: Die Gesamtstruktur des Gralis-Korpus

Arno Wonisch (Graz)

Das Gralis Text-Korpus

11. Der im Rahmen des Gralis-Portals eingerichtete Menüpunkt „Gralis-Korpus“ weist in seiner inneren Struktur drei Unterteilungen auf, von denen an dieser Stelle das Augenmerk dem „Gralis-Korpusarium“ gelten soll. Dieses stellt die konkreten Ergebnisse sämtlicher, im Jahre 2006 begonnener Aktivitäten auf dem Gebiet der Korpuslinguistik dar, die im April 2007 schließlich zur Inbetriebnahme und Eröffnung des Gralis-Korpus führten. Ein weiterer Blick in die Gliederung zeigt eine Unterteilung in zwei große Subkorpora – das Speech-Korpus und das Text-Korpus, wobei im Folgenden ein Einblick in die Nutzungsmöglichkeiten von Letzterem gegeben werden soll.

Das Gralis Text-Korpus besteht, wie bereits an anderer Stelle erwähnt, aus Texten, die in zumindest zwei sprachlichen Versionen (B und K, B und S, K und S) vorliegen, wobei das Augenmerk an dieser Stelle den Nutzungsmöglichkeiten des Korpus gelten soll.

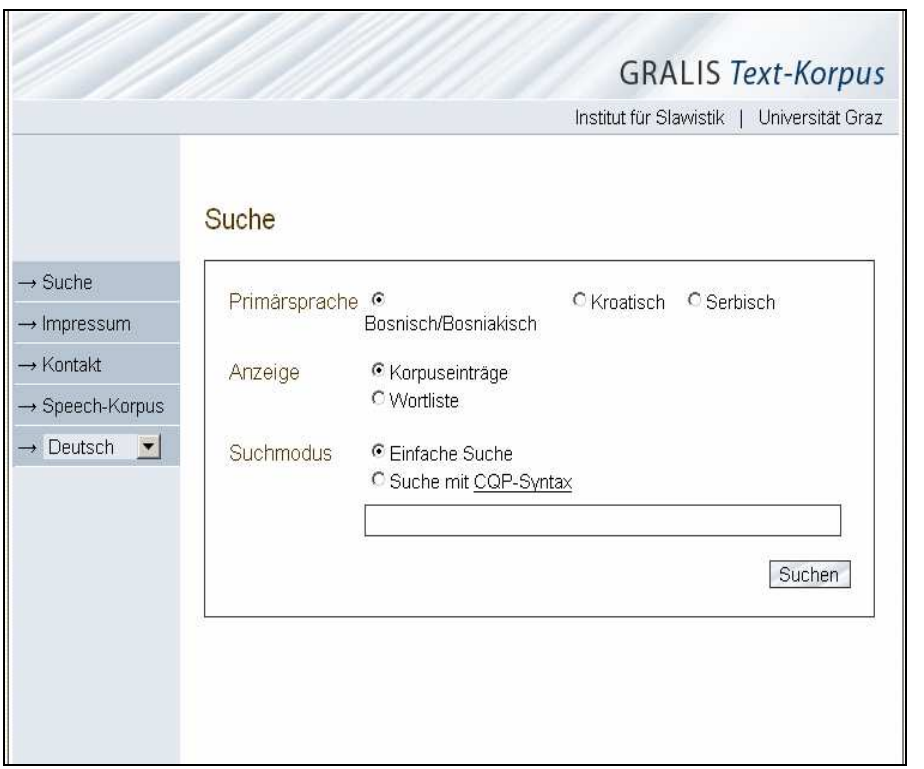


Abb. 9: Suchinterface des Gralis Text-Korpus (liegt auf Deutsch und BKS vor)

E i n f a c h e S u c h e . Ein mit dem Befehl „Einfacher Suche“ erhaltenes Suchergebnis bietet eine Gliederung nach den drei Sprachen entsprechend den internationalen Abkürzungen BS (bosnisch), HR (kroatisch) und

SR (serbisch), wobei die gewählte Primärsprache bei der Darstellung der Findstellen stets an erster Stelle erscheint. Das gesuchte Wort (bzw. die gesuchte Wortform) wird dabei abwechselnd mit gelber und blauer Farbe unterlegt und ausgewiesen. Oberhalb der Texte befindet sich der jeweilige Titel jenes Dokumentes (Roman, Zeitung u. Ä.), in dem der Suchbegriff gefunden wurde. Sollte es sich dabei um eine Internetquelle handeln, kann mittels Pfeil auf der rechten Seite des Titels die Linkadresse aufgerufen werden. Der in Abb. 9 dargestellte Auszug einer Eingabe des Sucheintrages *dom* stellt das Resultat des im Menüpunkt „Anzeige“ gewählten Befehles „Korpuseinträge“ dar, bei dem die aufgefundenen Wörter („Tokens“) in ihrer natürlichen Umgebung innerhalb eines Satzes erscheinen.

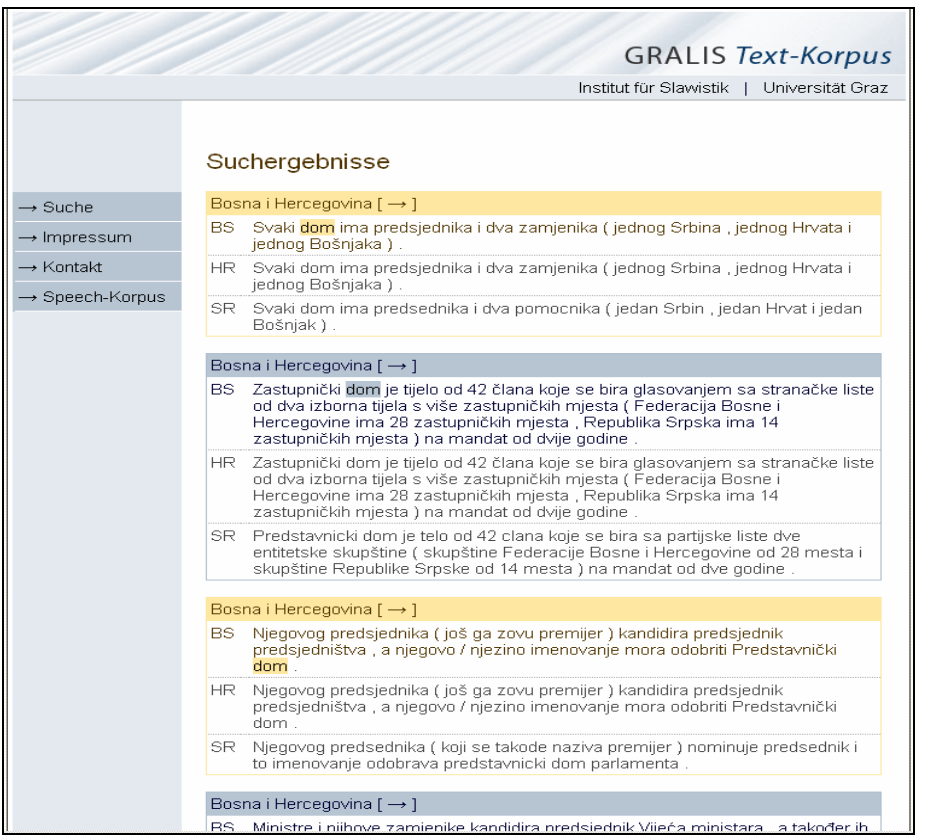


Abb. 10.: Ergebnis einer Abfrage im Gralis Text-Korpus (BKS) mit „HR“ als Primärsprache und Suchabfrage von *dom* (Heim, Haus)

Eine weitere Suchmöglichkeit liegt im Aufruf einer Frequenzliste, die die absolute Vorkommenshäufigkeit eines Suchbegriffes in der gewählten Primärsprache anzeigt, wozu der Befehl „Wortliste“ zu aktivieren ist.

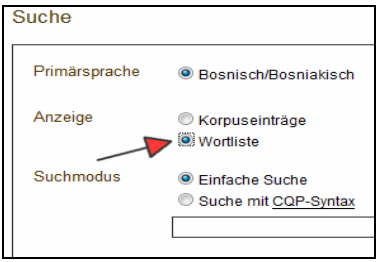


Abb. 11: Auswahl der Funktion der Frequenzliste

Nunmehr erscheint der Suchbegriff außerhalb seiner natürlichen Umgebung und mit dem Hinweis auf sein absolutes Vorkommen innerhalb des gesamten Korpus.

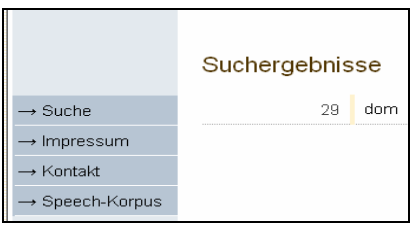


Abb. 12: Ergebnisse Frequenzliste für Sucheintrag *dom*

Die mithilfe einer einfachen Suche durchführbaren Nutzungsmöglichkeiten des Gralis Text-Korpus betreffen jedoch nicht nur ausschließlich die in Wörterbüchern verzeichneten Lemmata von Lexemen, sondern lassen auch gewisse morphologische und derivative Analysen zu. Dies betrifft etwa Suchabfragen mit den Symbolen **.** und *****, die in der hier dargestellten Abfolge im Falle von *dom.** sämtliche Tokens darstellen, die mit den Graphemen **d**, **o** und **m** beginnen.

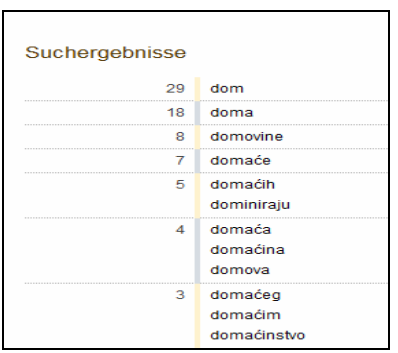


Abb. 13: Ergebnisse Frequenzliste für Sucheintrag *dom.**

Gleiches gilt auch für Abfragen, bei denen die Endung eines Wortes festgelegt wird und die Untersuchungen Morpheme im Wortinneren oder am Wortanfang gelten sollen. So etwa ergibt eine Suchabfrage der Zeichenfolge .**stvo* folgende Ergebnisse:

Abb. 14: Ergebnisse Frequenzliste für Sucheintrag .*stvo

S u c h e m i t C Q P - S y n t a x . Durch ein Aktivieren des Befehls „Suche mit CQP-Syntax“ im Menüpunkt „Suchmodus“ wird die als CQP bezeichnete Abfragesyntax der IMS Corpus Workbench aktiviert, die die technischen Möglichkeiten der im Rahmen der am Institut für maschinelle Sprachverarbeitung in Stuttgart entwickelten IMS Corpus Workbench (CWB)[1] ausschöpft. Auf diese Weise bieten sich den BenutzerInnen weitaus umfassendere Abfragemöglichkeiten, wobei sich jede Suchanfrage aus einem regulären Ausdruck (regular expression – Zeichenkette) in Kombination mit Metazeichen (attributive expressions – z. B. \, **!**, |, **{}**, **[]**) und einem Semikolon am Ende der Abfrage zusammensetzt. Eine Suchabfrage mit der CQP-Syntax eignet sich zur Durchführung jeglicher Art von linguistischen Analysen So zeigt eine Suchabfrage des Typs **"[hH]vala";** als Ergebnis all jene Korpuseinträge, in denen das Lexem *hvala* ungeachtet eines kleinen oder großen Anfangsbuchstabens vorkommt. Eine Zeichenfolge des Typs **".*t" "ć.*";** bildet etwa als Ergebnisse all jene Fälle ab, in denen das Graphem **t** am Ende eines Wortes und **ć** zu Beginn des darauf folgenden steht (*bit ću*). Gibt man **"da" [] {0,3} "se"** ein, so erhält man all jene Nennungen des Lexems *da*, auf das innerhalb der nächsten drei Wörter *se* folgt. Eine Abfrage von **"u" []*"u" []*"u"** führt zur Darstellung der kleinsten Segmente (im Falle des Gralis Text-Korpus sind dies Sätze), in denen dreimal die Präposition *u* auftritt. Durch die Eingabe von **"po.*" [pos="IN" | pos="PP"];** bzw. **"po.*" ([pos="IN"] | [pos="PP"]);** bzw. **"po.*" [pos="IN|PP"];** erhält man als Suchergebnis all jene im Korpus enthaltenen Tokens, die mit **po-** beginnen und auf die entweder eine Präposition oder ein Personalpronomen folgt. Gleiches gilt auch für den Befehl **"po.*" []{0,10} [pos="IN" | pos="PP"];**, doch können in diesem Fall bis zu zehn Wörter zwischen dem Morphem **po-** am Wortanfang und einer nachfolgenden Präposition oder einem Personalprono-

[1] http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/

men liegen. Bei einer Eingabe von **"po.*" [word!="\."]{0,10} [pos="IN" | pos="PP"];** entsprechen die Suchkriterien denjenigen von zuvor genanntem Beispiel, doch liegt die Bedingung vor, dass zwischen **po-** und einer Präposition bzw. einem Personalpronomen kein Punkt verzeichnet sein darf. Eine Suche nach bestimmten Wortarten kann mit Eingaben des Typs **[pos="JJ.*"] [pos="N.*"] "and|or" [pos="N.*"];** erfolgen, wobei hier die Abfolge Adjektiv – Nomen – Konjunktion – Nomen dargestellt werden soll. Abschließend sei für die zahlreichen Möglichkeiten von Suchabfragen mit der CQP-Syntax die Eingabe **[pos="N.*"] "sam|je" [pos="V.*" & word=".*io"];** dargestellt, bei der seitens der anfragenden Person folgende Informationen an das Korpus gerichtet werden: Nomen, gefolgt von entweder *sam* oder *je*, woraufhin ein auf **-io** endendes Verb steht.

Durch das Vorliegen der beiden Optionen „Einfache Suche“ und „Suche mit CQP-Syntax“ bieten sich den BenutzerInnen des Gralis Text-Korpus zwei den jeweiligen individuellen Erfordernissen entsprechende Nutzungsmöglichkeiten. Liegt einer Suchanfrage das Bedürfnis nach einem schnellen Abrufen von Informationen zu Grunde, empfiehlt sich eine einfache Suche, mit der Angaben zur statistischen Häufigkeit von Lexemen aber auch Analysen zu (primär) Morphologie oder Derivation durchgeführt werden können. Sollen jedoch umfassende und fundierte Untersuchungen zu Syntax, Semantik, Pragmatik, Textlinguistik usw. der Sprachen bosnisch/bosniakisch, kroatisch und serbisch vorgenommen werden, stellt die CQP-Syntax für diese Zwecke ein adäquates, hilfreiches und multimodal nutzbares Instrument dar.

Abschließend sei angemerkt, dass das Gralis Text-Korpus mit Jänner 2007 insgesamt knapp zwei Millionen Tokens enthielt, wobei sich die Zahl der Korpustexte beinahe im Wochentakt erhöht. Zum Zeitpunkt der Drucklegung vorliegender Publikation verfügte das Korpus noch über keine morphosyntaktische Annotation, doch soll im Zuge des nächsten Entwicklungsschrittes eine Erweiterung des Gralis Text-Korpus durch ein lemmatisiertes Subkorpus aus repräsentativen Texten erfolgen.

Sandra Forić (Graz)

Das Gralis Speech-Korpus

12. Das Gralis Speech-Korpus (http://www-gewi.uni-graz.at/gralis/) stellt eine systematische Sammlung ausgewählter Texte und von Audiomaterial für das Bosnische/Bosniakische, Kroatische und Serbische dar und setzt sich aus den Subkorpora Fix-Korpus, Wort-Korpus und Frei-Korpus sowie aus den Applikationen Akzentarium, Transkriptarium und Suprasegmentarium zusammen. Das Korpus-Interface kann in den Sprachen bosnisch/bosniakisch, kroatisch und serbisch abgerufen werden.

Die Arbeit im Gralis Speech-Korpus erfolgt in drei Phasen: 1. Aufnahme von Personen und Ausfüllen einer individuellen Aufnahmeevidenz, 2. Bearbeitung und Einfügung des aufgenommenen Materials und 3. Abfrage und Suche.

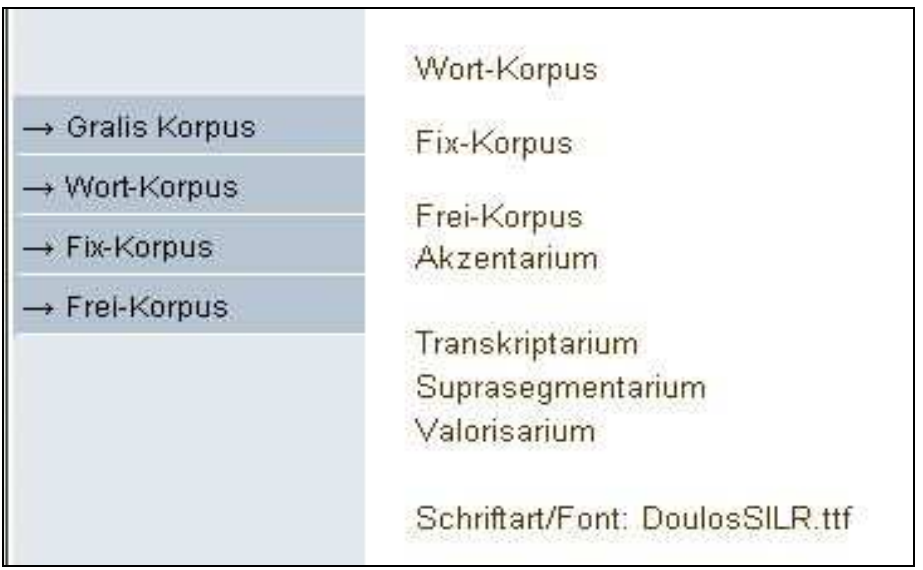


Abb. 15: Die Struktur des Gralis Speech-Korpus

Das Gralis Speech-Korpus besteht aus Texten in einer oder mehreren der genannten Sprachen aus sämtlichen Lebensbereichen, Genres und Stilen – von literarisch-künstlerisch über wissenschaftlich bis hin zu Lehrbüchern mit statistischen Angaben –, die ergänzend dazu auch über Audiofiles verfügen, die als Hilfe beim Erlernen von Sprachen und für phonetische Analysen genutzt werden können.

Abb. 16: Darstellung eines Suchergebnisses

Die Bearbeitung der Aufnahmen erfolgt in der Rubrik „Audiomaterial“, in der an erster Stelle das zuletzt eingestellte Audiofile mit sämtlichen Angaben und der Aufnahmeevidenz steht. Darunter befinden sich alle Informationen zur aufgenommenen Person, die aus Gründen der Datenanonymität allesamt unter einer Chiffre dargestellt werden. Alle eingegebenen Daten können zu jedem beliebigen Zeitpunkt bearbeitet und gelöscht werden, wobei jedes Login mit einer Datumsangabe versehen ist.

Unter „Status“ erscheint die Anzahl der Aufnahmen, die von einer mitarbeitenden Person angefertigt wurden. Man sieht zudem auch, welche Aufnahmen noch nicht fertig bearbeitet wurden und welche Angaben gemäß Aufnahmeevidenz noch einzutragen sind.

Abb. 17: Die Menüführung der Aufnahmeevidenz

Ein Verzeichnis sämtlicher Angaben ist über das Fix- und über das Wort-Korpus abrufbar, wobei man bei Anwahl des Fix-Korpus eine Liste der bestehenden Texte *Jutro*, *Na obali* und *Moja prijateljica* erhält, die auf Satzebene segmentiert wurden, während das Wort-Korpus eine Liste mit 99 Wörtern enthält.

Die Segmente werden in den Formaten wav und mp3 in das Korpus eingespeist, wobei mp3 für einen schnelleren Download dient und wav alle für eine Spektralanalyse erforderliche Frequenzhöhen beibehält.

In unten stehender Tabelle sind alle Angaben zu den einzelnen zu analysierenden Segmenten beinhaltet. Mit einem Klick auf den Eintrag „Šifra“ öffnet sich die Aufnahmeevidenz, in der die Angabe der Stadt den Aufnahmeort, die dreistellige Nummer die fortlaufende Zahl der aufgenommenen Person und der Kleinbuchstabe am Ende schließlich die jeweilige Muttersprache bezeichnet. Befindet sich hinter der Sprachangabe der Buchstabe **w**, so ist dies ein Hinweis darauf, dass die Aufnahme Teil des Wort-Korpus ist. Oberhalb des Verzeichnisses befindet sich die Seitenanzahl, wobei die zuletzt in das Korpus eingefügte Audiodatei stets an oberster Stelle der ersten Seite erscheint.

Spisak/Lista unesenih podataka

Wort- und Fix-Korpus

Dobijeni rezultat: 1 - 20 od 105 1 2 3 4 5 6

Ev.	P.	Šifra	Val.	Datum	Snimak/snimku un(ij)eti									Evidenciju preraditi
					Puni snimak/ Puna snimka		Segmentovani snimak/ Segmentirana snimka					Uvod		
					wav	mp3	wav	mp3	Akzent	Transk.	Int.	wav	mp3	
225	190	graz_011b		04.05.2007 17:26	0	0	0/0	0/0	0/0	0/0	0/0	0/4	0/4	
224	189	graz_010b		04.05.2007 17:23	0	0	0/0	0/0	0/0	0/0	0/0	0/4	0/4	
223	188	graz_009b		04.05.2007 17:02	0	0	0/0	0/0	0/0	0/0	0/0	0/4	0/4	
169	143	kapela_036k		12.04.2007 00:49	0	0	0/18	0/18	0/18	18/18	0/18	0/4	0/4	
168	143	kapela_035k		12.04.2007 00:48	0	0	0/18	0/18	0/18	0/18	0/18	0/4	0/4	
167	143	kapela_034k		12.04.2007 00:48	0	0	0/18	0/18	0/18	0/18	0/18	0/4	0/4	
166	142	kapela_033k		12.04.2007 00:48	0	0	0/18	0/18	0/18	0/18	0/18	0/4	0/4	
165	142	kapela_032k		12.04.2007 00:44	0	0	0/18	0/18	0/18	0/18	0/18	0/4	0/4	
164	141	kapela_031k		12.04.2007 00:43	0	0	0/18	0/18	0/18	0/18	0/18	0/4	0/4	
163	141	kapela_030k		12.04.2007 00:43	0	0	0/18	0/18	0/18	0/18	0/18	0/4	0/4	
162	141	kapela_029k		12.04.2007 00:41	0	0	0/18	0/18	0/18	0/18	0/18	0/4	0/4	
161	141	kapela_028k		12.04.2007 00:40	0	0	0/18	0/18	0/18	0/18	0/18	0/4	0/4	
160	140	kapela_027k		12.04.2007 00:38	0	0	0/18	0/18	0/18	0/18	0/18	0/4	0/4	
157	22	zadar_001kw		04.04.2007 19:00	0	0	0/99	0/99	0/99	0/99	0/99	0/4	0/4	
156	90	zenica_001kw		04.04.2007 19:00	0	0	0/99	99/99	1/99	0/99	0/99	0/4	0/4	
158	88	osijek_001kw		04.04.2007 18:59	0	0	0/99	99/99	0/99	0/99	0/99	0/4	0/4	
159	28	bjelovar_009kw		04.04.2007 18:59	0	0	0/99	99/99	0/99	0/99	0/99	0/4	0/4	
92	92	zenica_001kb		03.04.2007 18:40	0	0	0/12	0/12	0/12	0/12	0/12	0/4	0/4	
91	91	zenica_001ka		03.04.2007 18:40	0	0	0/13	0/13	0/13	0/13	0/13	0/4	0/4	
82	82	sarajevo_001ba		03.04.2007 18:39	0	0	0/13	0/13	0/13	0/13	0/13	0/4	0/4	

Abb. 18: Übersicht über eingetragene Aufnahmeevidenzen und Audiodateien

Die Einführung bzw. Vorstellung der aufgenommenen Person besteht aus der personenbezogenen Chiffre (1), dem Alter (2), Geburtsort (3) und der Muttersprache (4) der befragten Person.

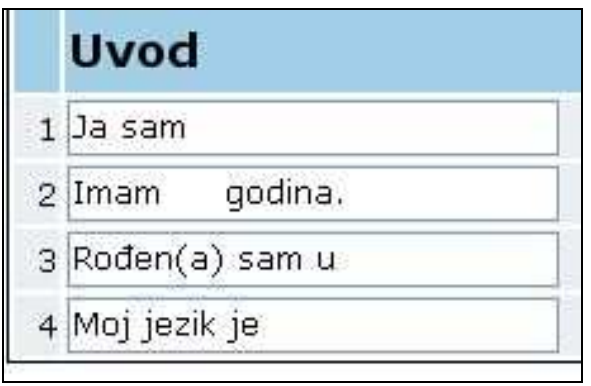


Abb. 19: Die Einführung (zu Beginn jeder Aufnahme)

Das bereits erwähnte Frei-Korpus stellt im Gegensatz zum Fix- und Wort-Korpus einen integralen Bestandteil des Gralis Text-Korpus dar, das bislang die Aufnahme zu einer Person umfasst (Ujak 2002).

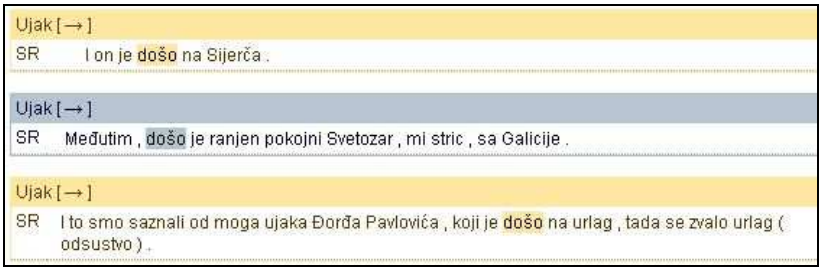


Abb. 20: Auszug aus dem Gralis Frei-Korpus

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.

Abb. 21: Bearbeitungsstatus von Aufnahmen (Audioformate, Akzentuierung, Transkription, Intonation)

A k z e n t . Jede mitarbeitende Person legt mittels Audio- oder Spektralanalyse die Akzente der einzelnen Lexeme fest, wobei nach Durchführung einer auf Gehör basierenden Akzentuierung die Dateien unmittelbar nach Abschluss der Arbeiten ins Valorisarium überführt werden, in dem Fachleute eine Beurteilung der Akzente vornehmen. Stimmen die Meinungen dreier ExpertInnen für Akzentologie überein, wird das Wort bzw. eine komplett bearbeitete Datei für die weitere Analyse freigegeben. Sollten die Ansichten divergieren, obliegt die letzte Entscheidung dem Projektleiter, wobei generell darauf hingewiesen sei, dass dieser Arbeitsschritt für alle Beteiligten ein oftmaliges Abhören des Audiomaterials erforderlich macht.

graz_005k

Jutro

Evidenz graz_005k ist unter Rezension.

	mp3	Segmentovani snimak/ Segmentirana snimka	Akzent 18 / 18
1		Jutros su me vrlo rano ptice probudile.	jûtros su me vřlo rãno pt̂ice probúdile
2		Sunce je tek bilo izašlo iza brda.	sûnce je tȅk bílo ìzašlo iza brda
3		U daljini se prostirala velika šuma.	u daljìni se pròstirala vȅlika šûma
4		U njoj su rasli borovi, jele, breze i omorike.	U njôj su rásli bórovi, jéle, bréze i omòrike.

Abb. 22: Akzentuierung

Die V a l o r i s i e r u n g stellt die finale Bewertung seitens ausgewiesener Fachleute auf dem Gebiet der Akzentologie dar. Mit ihr endet die Bearbeitung von Aufnahmen im Gralis Speech-Korpus, sodass danach mit der Spektralanalyse begonnen werden kann. Sämtliche von den Fachleuten vorgenommenen Akzentuierungen scheinen dabei neben dem jeweiligen Satz auf. Um einen Text valorisieren zu können, ist in der Rubrik „Audiomaterial“ die entsprechende, individuelle Chiffre einzugeben. In der für die Valorisierung vorgesehen Rubrik befinden sich in der ersten Reihe die akzentuierten Sätze, die nun von den Fachleuten auf ihre Richtigkeit hin zu überprüfen sind. Sämtliche Einträge der ExpertInnen scheinen sodann auf der rechten Seite des überprüften Satzes auf, wobei durch einen Klick auf das Ordnersymbol eventuell falsch gesetzte Akzente eingesehen werden können.

	mp3	Rečenica/Akzent/Transkripcija/Intonacija			Experten				
1		Jutros su me vrlo rano ptice probudile.							
		A jûtros su me vřlo rãno pt̂ice probúdile			1	!	Kozomara:	jûtrōs	06.11.2007 15:08
		T	-		1	✓	Kozomara:		06.11.2007 15:08
		I	-		1	✓	Kozomara:		06.11.2007 15:08
2		Sunce je tek bilo izašlo iza brda.							
		A sûnce je tȅk bílo ìzašlo iza brda	-		1	✓	Kozomara:		06.11.2007 15:09
		T			1	!	Kozomara:		06.11.2007 15:09
		I			1	!	Kozomara:		06.11.2007 15:09

Abb. 23: Valorisierung

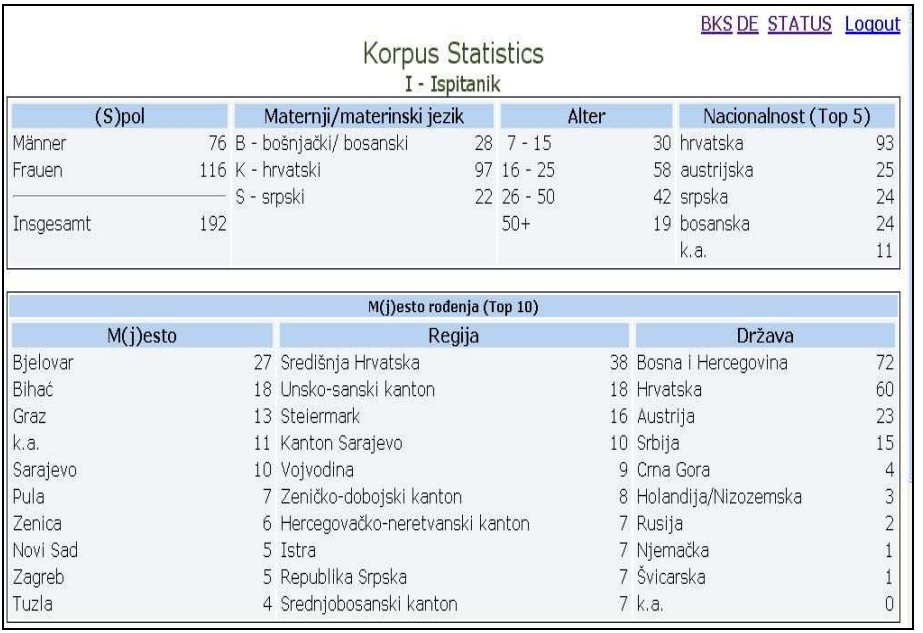
I n t o n a t i o n . Die Bestimmung der Intonation läuft in zwei Phasen ab. Mittels Audioanalyse wird in einem ersten Arbeitsschritt der Tonverlauf festgelegt, wobei zwischen steigender, gleichbleibender sowie fallender Satzmelodie unterschieden wird und Sprechpausen markiert werden. Daraufhin wird das bereits beschriebene Programmskript Prosogramm geöffnet, von der Darstellung ein Sreenshot angefertigt und in das Korpus eingefügt.

Abb. 24: Intonation

T r a n s k r i p t i o n . Die technische Vorgangsweise zur Niederschrift der Transkription entspricht jenen zur Bestimmung der Akzente und der Satzintonation. Als primäres Alphabet dient dazu die so genannte Gralis-Transkription, die eigens für die Bedürfnisse der Sprachen bosnisch/bosniakisch, kroatisch und serbisch entwickelt wurde und später in international übliche Transkriptionsalphabete wie SAMPA oder IPA überführt werden kann.

Abb. 25: Transkription

Zur Erlangung eines statistischen Überblickes über die Zahl der aufgenommenen Personen und den aktuellen Umfang des Gralis Speech-Korpus besteht die Möglichkeit, die so genannte „Korpus Statistics“ abzurufen (siehe unten stehende Abbildung) die Parameter wie Geschlecht, Muttersprache, Alter, Nationalität, Geburtsort, -region und -land darstellt.



BKS DE STATUS Logout

Korpus Statistics

I - Ispitanik

(S)pol		Maternji/materinski jezik		Alter		Nacionalnost (Top 5)	
Männer	76	B - bošnjački/ bosanski	28	7 - 15	30	hrvatska	93
Frauen	116	K - hrvatski	97	16 - 25	58	austrijska	25
		S - srpski	22	26 - 50	42	srpska	24
Insgesamt	192			50+	19	bosanska	24
						k.a.	11

M(j)esto rođenja (Top 10)

M(j)esto		Regija		Država	
Bjelovar	27	Središnja Hrvatska	38	Bosna i Hercegovina	72
Bihać	18	Unsko-sanski kanton	18	Hrvatska	60
Graz	13	Steiermark	16	Austrija	23
k.a.	11	Kanton Sarajevo	10	Srbija	15
Sarajevo	10	Vojvodina	9	Crna Gora	4
Pula	7	Zeničko-dobojski kanton	8	Holandija/Nizozemska	3
Zenica	6	Hercegovačko-neretvanski kanton	7	Rusija	2
Novi Sad	5	Istra	7	Njemačka	1
Zagreb	5	Republika Srpska	7	Švicarska	1
Tuzla	4	Srednjobosanski kanton	7	k.a.	0

Abb. 26: Korpusstatistiken

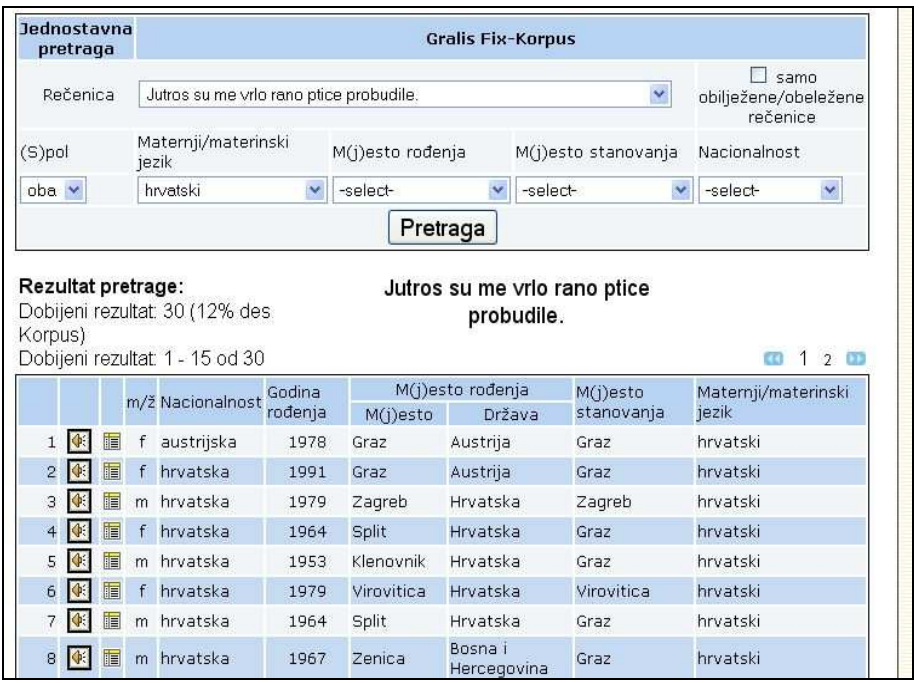
Von besonderer Bedeutung für die Bestimmung von Akzenten erweist sich das Akzentarium, in dem durch Eingabe eines Suchbegriffes die jeweiligen Akzentuierungen in den Sprachen bosnisch/bosniakisch, kroatisch, serbisch und serbokroatisch angezeigt werden. Als Quelle für die einzelnen Lexeme dienen dabei 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.

Abb. 27: Akzentarium

Suchabfragen im Gralis Speech-Korpus. Das Gralis Speech-Korpus stellt eine Sammlung von Texten und Aufnahmen dar, die auf einfache Weise durchsucht werden können: 1. mittels Wahl des ent-

sprechenden Subkorpus (Text- oder Speech-Korpus) und 2. durch die Wahl von Texten (im Fix-Korpus, Wort-Korpus oder Frei-Korpus).

In weiterer Folge kann die Auswahl der Sprache, des Geburtsortes, der Nationalität, des Geschlechts und des Segmentes (d. h. eines Satzes oder Wortes) vorgenommen werden, das zu hören gewünscht wird. Das Korpus bietet eine Reihe an Kombinationsmöglichkeiten im Rahmen der personenbezogenen Angaben, die schließlich gemäß den gewählten Parametern zur Darstellung des gesuchten Audiomaterials führen.



			m/ž	Nacionalnost	Godina rođenja	M(j)esto rođenja		M(j)esto stanovanja	Maternji/materinski jezik
						M(j)esto	Država		
1			f	austrijska	1978	Graz	Austrija	Graz	hrvatski
2			f	hrvatska	1991	Graz	Austrija	Graz	hrvatski
3			m	hrvatska	1979	Zagreb	Hrvatska	Zagreb	hrvatski
4			f	hrvatska	1964	Split	Hrvatska	Graz	hrvatski
5			m	hrvatska	1953	Klenovnik	Hrvatska	Graz	hrvatski
6			f	hrvatska	1979	Virovitica	Hrvatska	Virovitica	hrvatski
7			m	hrvatska	1964	Split	Hrvatska	Graz	hrvatski
8			m	hrvatska	1967	Zenica	Bosna i Hercegovina	Graz	hrvatski

Abb. 28: Darstellung von Suchergebnissen

Durch eine Anwahl des Lautsprecher-Symbols  kann die gewählte Audiodatei schließlich angehört werden. Am oberen Rand der Aufnahmeliste erscheint dazu die statistische Angabe (in absoluten Zahlen und prozentuell), wie oft der gesuchte Eintrag mit den entsprechenden Parametern insgesamt im Korpus enthalten ist.

Abb. 29: Anzahl der Einträge mit den gesuchten Parametern

Neben dem Lautsprechersymbol befindet sich eine Darstellung in Form eines Textdokumentes, mithilfe derer alle von einer Person gesprochenen Sätze angehört werden können.

		Dobijeni rezultat: 1 Tema: Jutro
1		Jutros su me vrlo rano ptice probudile.
2		Sunce je tek bilo izašlo iza brda.
3		U daljini se prostirala velika šuma.
4		U njoj su rasli borovi, jele, breze i omorike.
5		Na livadama oko šume pasle su ovce i krave.
6		Na travi je i dalje ležala rosa.
7		U obližnjem parku ptice su cvrkutale.
8		Grad se postepeno budio.
9		Ulice su postajale sve bučnije.
10		U dvorištu su se pojavila dva psa.
11		Oni su se veselo igrali, povremeno lajući.
12		Iza dvorišta su dopirali zvuci tramvaja.
13		Đaci su se spremali za odlazak u školu.
14		Neki od njih imali su na sebi džempere.
15		Trgovine su počele s radom.
16		Poštari su raznosili pisma.
17		Jutro je sve više prelazilo u dan.
18		Ja se spremam za rad.

Abb. 30: Option zum Anhören aller von einer Person gesprochenen Sätze

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.

Maja Midžić (Graz)

Die Aufnahmeevidenz des Gralis Speech-Korpus

13. Die Aufnahmeevidenz stellt einen integralen Bestandteil des Gralis Speech-Korpus (bestehend aus dem Fix-, Wort- und Frei-Korpus) dar und dient zur Verwaltung der Informationen zu den aufgenommenen Personen und den einzelnen Audiofiles. Der Evidenzeintrag bildet den ersten Arbeitsschritt unmittelbar nach der Aufnahme, wobei das erfasste Audiomaterial entweder aus Sätzen oder Wörtern besteht, für die die Evidenz des Wort- oder des Fix-Korpus auszufüllen sind. Handelt es sich um eine Aufnahme freier, spontaner Rede wie etwa um ein Interview, einen Monolog, Dialog, ein Gespräch u. Ä., werden die Metadaten zur/zum Sprechenden und das Audiofile mit der Evidenz des Frei-Korpus erfasst. Die Evidenz ist auf Deutsch und Bosnisch/Bosniakisch, Kroatisch und Serbisch (im Folgenden: BKS) abrufbar und besteht aus vier Teilen. Im ersten befinden sich Angaben zur befragten Person, d. h. derjenigen Person, die aufgenommen wurde, der zweite Teil beinhaltet Informationen zur Aufnahme selbst, der dritte Teil betrifft die Analyse, und im vierten Teil ist schließlich eine schriftliche Einverständniserklärung durch die befragte Person abzugeben. In den Teilen zwei und vier erfolgt eine Unterscheidung zwischen Fix-, Wort- und Freikorpus, während die beiden anderen für alle drei Subkorpora identisch sind. Um mit dem Ausfüllen der Evidenz beginnen zu können, ist es in einem ersten Arbeitsschritt erforderlich, für jede Person bzw. jede Aufnahme eine individuelle und unikale Chiffre festzulegen. Diese erhält man dadurch, indem zuerst der Ort, an dem die Aufnahme stattfindet, eingetragen wird, wobei die Schreibweise des Ortes derjenigen in der Originalsprache entspricht und ausschließlich in Kleinbuchstaben erfolgt. Besteht ein Ortsname aus zwei oder mehreren Wörtern, werden diese durch Unterstriche getrennt. Auf den Ortsnamen folgt eine dreistellige Zahl, die für jede Sprache mit dem Eintrag „001“ beginnt.

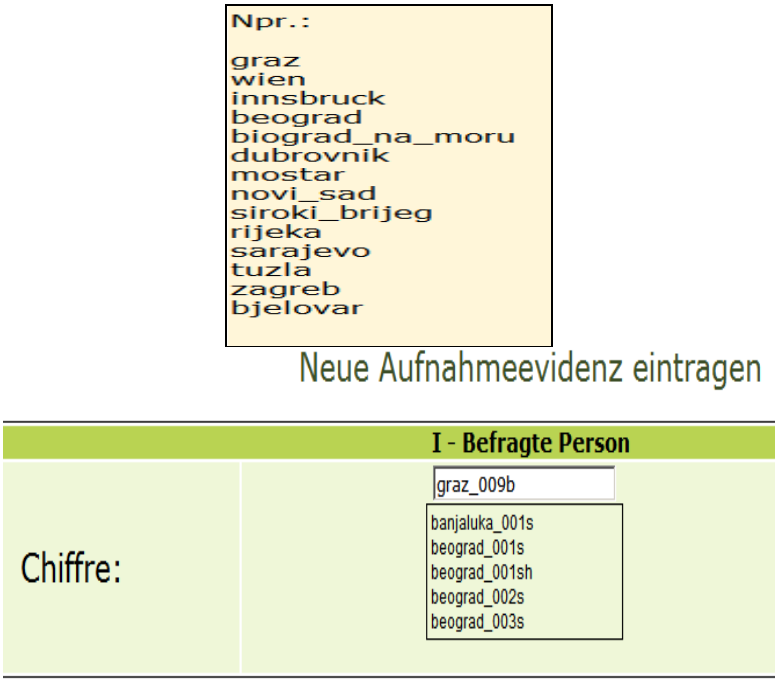


Abb. 31: Eintragen einer neuen Evidenz

Das Ende der Chiffre bildet die Abkürzung für die Muttersprache der aufgenommenen Person, wobei sich ein Verzeichnis der Kürzel wie auch aller anderen Parameter zur Festlegung der Chiffre auf der Startseite der Aufnahmeevidenz befindet.

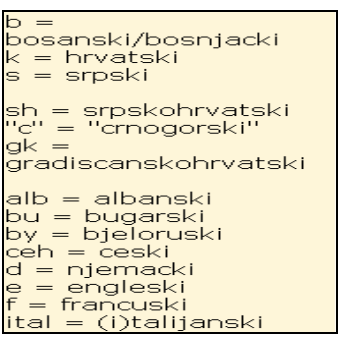


Abb. 32: Kürzel für die einzelnen Sprachen

Wird die Aufnahme in Graz durchgeführt und handelt es sich bei der aufgenommenen Person um die siebente mit der Muttersprache bosnisch, so lautet die Chiffre in diesem Fall graz_007b. Inklusive Hinzufügung des Alters sind auf diese Weise vor dem Verlesen des Textes (Sätze oder Wörter) folgende Metainformationen bekannt zu geben: „Ja sam graz_007b. Imam … godina/godine. Rođen(a) sam u … Moj jezik je bosanski. [Ich bin graz_007b.

Ich bin … Jahre alt. Ich wurde in … geboren. Meine Sprache ist bosnisch.] Die Zahl hängt davon ab, um die wievielte aufgenommene Person es sich mit dieser oder jener Muttersprache in der jeweiligen Stadt handelt. Wurde die Chiffre schließlich definiert, folgt als nächster Schritt das Ausfüllen der Evidenz, wobei dies sowohl von der aufnehmenden als auch von der aufgenommenen Person durchgeführt werden kann. Das erste Feld betrifft das Geschlecht (f/m), die Nationalität (bosnisch/bosniakisch, kroatisch, serbisch usw.), Religion (orthodox, katholisch, muslimisch u. a.) und das Geburtsjahr. Es folgen in der zweiten Rubrik der Geburtsort, die Region bzw. die politische Verwaltungseinheit, in der sich dieser befindet und abschließend der Staat, z. B.: Graz, Steiermark, Österreich. Der dritte Abschnitt der Evidenz beinhaltet Angaben zum Wohnort, gefolgt von der vierten Rubrik mit Angaben zum Beruf (Studentin/Student, Schülerin/Schüler, Angestellte(r) usw.), zu Arbeits- bzw. Ausbildungsstätte (Firma oder Universität, Fakultät, Institut, Schule; z. B.: Karl-Franzens-Universität Graz, Geisteswissenschaftliche Fakultät, Institut für Slawistik) zum Ort, an dem der Arbeit bzw. der Ausbildung nachgegangen wird und zum wissenschaftlichen Grad. Der nächste Eintrag definiert die Bildung (höhere, mittlere, Pflichtschulbildung, keine) und Ort sowie Zeit des Schul- bzw. Universitätsbesuches (z. B.: Zagreb, 1986–1994). Von wesentlicher Bedeutung sind die darauf folgenden Angaben zur Muttersprache der aufgenommenen Person, zum Dialekt (štokavisch, kajkavisch, čakavisch), zur regionalen Variante und zur Aussprache (ijekavisch, ekavisch und ikavisch) sowie zur Muttersprache der beiden Elternteile. Der vorletzte personenbezogene und überaus wichtige Eintrag betrifft die einzelnen Lebensmittelpunkte, gefolgt von den Fremdsprachenkenntnissen, die ebenso wie ein Wechsel des Wohnortes in hohem Maße auf die Ausformung der Sprache Einfluss nehmen können.

Im zweiten Teil der Evidenz wird das Thema der Aufnahme eingetragen, d. h. der Titel des Textes, der von einer aufgenommenen Person verlesen wird, woraufhin Ort und Datum der Aufnahme folgen. In der zweiten Reihe wird auf die Situation, in der eine Aufnahme entstand, hingewiesen (in einer Wohnung, auf der Straße, in einem öffentlichen Objekt u. a.). Die dritte Rubrik beinhaltet Angaben zum Aufnahmegerät, wobei zwischen einem speziellen Recorder zur Beibehaltung der hohen Frequenzen und Diktiergeräten der Marke Sony und Olympus unterschieden wird. Die letzte Zeile der Evidenz bilden Informationen zur genauen Dauer der Aufnahme und zu deren Audioformat (wma, wav, mp3, ogg, aac, m4a, cda u. a.).

Der dritte Teil der Evidenz umfasst Angaben zur Art der Analyse, wobei in einem ersten Teil definiert wird, ob es sich um eine Audio- oder Spektralanalyse handelt und in einem zweiten Teil Platz für eventuelle Anmerkungen zu aufgenommener Person oder Aufnahme vorgesehen sind.

Den vierten und letzten Teil der Evidenz bildet schließlich die schriftliche Einverständniserklärung, dass die Aufnahme für wissenschaftliche Zwe-

cke herangezogen werden darf und lautet wie folgt: „Slažem se da se ovaj snimak / ova snimka pod šifrom i s(a) navedenim podacima uključi u Gralis-Korpus (http:www-gewi.uni-graz.at/gralis/).“ [Ich bin einverstanden, dass diese Aufnahme unter einer Chiffrenummer und mit den getätigten Angaben in das Gralis-Korpus aufgenommen wird.] Es folgt die Eingabe von Aufnahmeort und -datum.

Die Evidenz für das Frei-Korpus unterscheidet sich von denen für das Fix- und Wort-Korpus dadurch, dass nach dem Thema, Ort und Datum auch die Art der Aufnahme einzugeben ist. Es kann sich dabei um einen Monolog, Dialog, ein Interview, eine Lesung, ein Gespräch, eine Diskussion oder um einen runden Tisch handeln. Danach folgt der Verweis zum funktionalen Stil (literarisch-künstlerisch, administrativ, publizistisch oder umgangssprachlich) und abschließend zum Medium (TV, Radio, Film, Skype u. a.). Im dritten Teil der Evidenz ist bei allen Subkorpora einzutragen, um welchen der drei es sich handelt (Wort-, Fix- oder Frei-Korpus).

Beim Ausfüllen der Evidenz bietet sich die Möglichkeit des Hinzufügens neuer Tabelleneinträge, wozu der Befehl „Novo m(j)esto ili novi jezik un(ij)eti“ [Neuen Ort oder neue Sprache einfügen] anzuwählen ist, mit dem folgende Felder ergänzt werden können: Analyse, Apparat, Beruf, Format, Dialekt, Religion, Nationalität, Ort (für Geburts- und Wohnort), regionale Variante, Situation (der Aufnahme), Sprache, Staat, Thema und (akademischer) Titel. Aktiviert man nun diesen Befehl zur Ergänzung der Einträge, öffnet sich eine graphisch unterschiedlich gestaltete Tabelle, in der sich auf der linken Seite die bereits bestehenden Einträge und rechts leere Felder befinden, in die der gewünschte Begriff hinzugefügt werden kann. Die Evidenz kann auch bearbeitet werden, sodass sämtliche Einträge zu jedem Zeitpunkt abgeändert werden können.

Angesichts dessen, dass die Angaben und Einträge für alle drei Subkorpora nahezu identisch sind und ein- und dieselbe Person mehrere Texte verlesen kann, ist es möglich, eine bereits bestehende Evidenz zu duplizieren, wodurch kein erneutes Eingeben gleicher Tabelleninhalte erforderlich ist. In weiterer Folge können zu ändernde Angaben in die bereits ausgefüllte Evidenz eingegeben werden. Gleiches gilt auch für zwei Personen, deren Angaben sich in hohem Maße gleichen.

Abschließend sei ein Beispiel einer vollständig ausgefüllten Evidenz dargestellt:

BKS DE STATUS Logout

Aufnahmeevidenz:

I - Befragte Person					
Chiffre:	bihac_012bw		Geschlecht: m		
Nationalität:	bošnjačka	Konfession:	muslimanska	Geburtsjahr:	1949
Geburtsort	Bihać	Unsko-sanski kanton	Bosna i Hercegovina		
Wohnort	Bihać	Unsko-sanski kanton	Bosna i Hercegovina		
Beruf	profesor	Fakultät, Schule, Institution, Firma. Matematički fakultet		Ort : Bihać	Titel: profesor
Bildung:	visoko				
Ort des Schulbesuchs bzw. Studiums	1. Ort: Banja Luka von 1975 bis 1977 2. Ort: Rijeka von 1968 bis 1975 3. Ort: Bihać von 1956 bis 1968				
Muttersprache	monolingual: bosanski	polylingual:	Dialekt: štokavski 100 % kajkavski 0 % čakavski 0 %	Idiom: --	Regionale Variante: - -
Aussprache	Ijekavski				
Muttersprache der Eltern	Mutter:	bosanski		Vater:	bosanski
Lebens- mittelpunkte	1. Ort: Bihać von 1977 bis 2007 2. Ort: Banja Luka von 1975 bis 1977 3. Ort: Rijeka von 1968 bis 1975 4. Ort: Bihać von 1949 bis 1968				
Fremdsprache	1. nemački/njemački				
II - Aufnahme					
Thema	100 rijeci	Aufnahmeort:	Bihać	Datum:	12.07.2007
Situation	u stanu				
Aufnahmeapparat	Typ:	aparat za diktiranje	Apparat:	Olympus	
Länge der Aufnahme: Stunde, Minute, Sekunde:	00:01:30	Format:	wma	Aufnehmende Person:	Maja Midzic
III - Korpus					
Analyse:	slušna		Korpus:	Wort-Korpus	
Anmerkungen:					

Abb. 33: Beispiel einer abgeschlossenen Aufnahmeevidenz

Branko Tošović (Graz)

Das Gralis-Akzentarium

14. Der Akzent ist in einigen Sprachen an eine gewisse Position innerhalb eines Wortes gekoppelt. Einige Sprachen besitzen nur einen einzigen Akzent, der an eine fixe Silbe gebunden sein kann, nämlich an die erste, zweite, drittletzte (die sog. Antepenultima), vorletzte (die sog. Penultima) und an die letzte Silbe (die sog. Ultima). Dabei kann der der Akzent frei und beweglich sein und sich auf allen Silben oder auf der Mehrzahl der Silben befinden. Andere Sprachen verfügen über zwei, drei oder mehrere Akzente. Zur Gruppe der Sprachen mit fixem Akzent gehört z. B. das Tschechische (Betonung auf der 1. Silbe), das Polnische (Betonung auf der Penultima) und das Mazedonische (Betonung auf der Antepenultima). Ein typisches Beispiel für einen freien und beweglichen Akzent stellen das Russische und das Slowenische dar, in denen der Akzent auf jeder beliebigen Silbe liegen kann.

Die schwierigste Situation trifft man im Bosnischen/Bosniakischen, Kroatischen und Serbischen an, wobei dies gleich mehrere Gründe hat: 1. Der Akzent ist nicht fix an eine Silbe gebunden, 2. er kann innerhalb von Paradigmen sehr oft beweglich sein (in der Deklination oder in der Konjugation kann sich seine Qualität oder Quantität ändern; manchmal auch ein Vorziehen auf eine Präposition erfolgen), 3. es liegen vier Akzente (lang steigend, kurz steigend, lang fallend, kurz fallend) und eine posttonale Länge vor, und 4. es gibt nur einige wenige Regeln (fallende Akzente können niemals außerhalb der ersten Silbe stehen, einsilbige Wörter verfügen ausschließlich über fallende Akzente, der Akzent kann auf allen Silben, ausgenommen auf der letzten zum Liegen kommen). Einige štokavische Dialekte des BKS (z. B. zetsko-južnosandžački govori – die Mundarten von Zeta und des Südsandžak). Ein Drei-Akzent-System ist für die slowenische Sprache und für čakavische und kajkavische Dialekte charakteristisch. Innerhalb all dieser Strukturen ist eine Orientierung oftmals nur sehr schwer möglich, wobei dies nicht nur für all jene, gilt, die BKS als Fremdsprache lernen, sondern auch für Personen mit BKS als Muttersprache zutrifft. Allein die Tatsache, dass einzig philologisch ausgebildete Sprechende die Akzente korrekt erfassen und verzeichnen können, zeugt von der Schwierigkeit dieser Thematik, die beim Erlernen des BKS das wohl größte Problem darstellt.

Aus eben diesen Gründen wurde im Rahmen des Projektes „Die Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen“ der Versuch unternommen, ein Online-Programm zu entwickeln, das für all jene Personen eine Hilfe darstellen möge, die die Akzente des BKS erlernen oder diese untersuchen möchten. Das somit im Rahmen des Gralis Speech-Korpus erarbeitete Gralis-Akzentarium stellt ein Programm für die Erfassung, Bearbeitung und Analyse der Akzente des Bosnischen/Bosniakischen, Kroatischen und Serbischen. Das Akzentarium wurde

auf Basis einer MySQL-Datenbank entwickelt und besteht aus zwei Teilen – dem so genannten Akzentor (für den automatischen Eintrag von Akzenten) und dem Interface für die Auffindung und Analyse bestimmter Akzente.

Der Akzentor dient für eine leichte und effiziente Akzentuierung von Wörtern und Wortformen. Durch ein Aufrufen dieser Funktion öffnet sich eine Maske, an deren oberen Rand sich der Titel des Textes befindet (z. B. Test_001k), aus dem ein zu bearbeitender Satz gewählt wird (z. B. *Jutros su me vrlo rano ptice probudile*.). Unterhalb dieses Menüpunktes folgt jener Teil, in dem sie Akzente eingetragen werden.

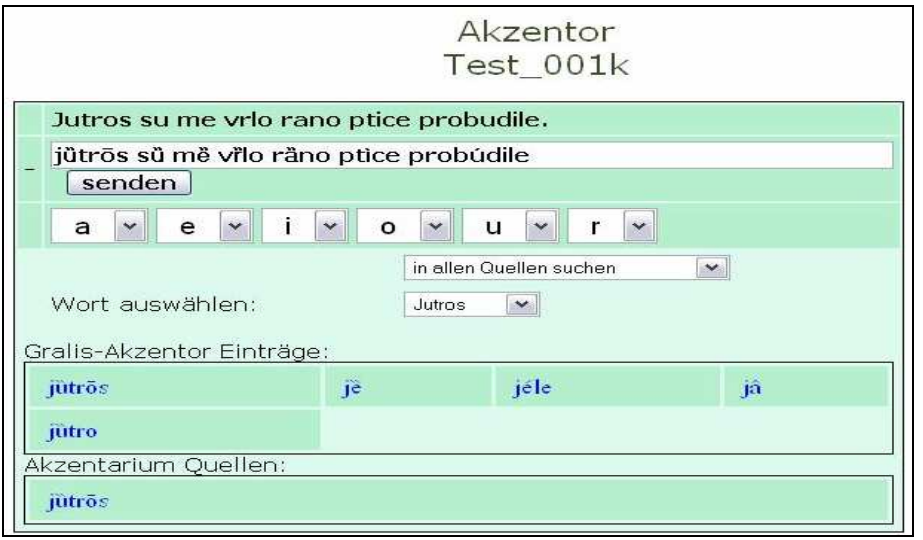


Abb. 34: Der Gralis-Akzentor

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.

á	à	â	a̋	ā
é	è	ê	e̋	ē
í	ì	î	i̋	ī
ó	ò	ô	ő	ō
ú	ù	û	ű	ū
ŕ	ṙ	r̂	r̋	ṛ

Ab. 35: Die Akzentsymbole im Gralis-Akzentarium

Z. B.:

rúka sèstra mâjka nȍga rúkā

Die einzige Ausnahme in der graphischen Darstellung stellt der kurz fallende Akzent für **r** dar, für den kein entsprechendes Zeichen im Unicode-Schriftsatz vorhanden ist, weshalb an dessen Stelle **ṙ** verwendet wird. Die Kodierungstabelle stellt sich wie folgt dar:

Latin-1		Latin-1		Latin-1		**Latin Extended-B**		**Latin Extended-A**	
á	225	à	224	â	226	a̋	513	ā	257
é	233	è	232	ê	234	e̋	517	ē	275
í	237	ì	236	î	238	i̋	521	ī	299
ó	243	ò	242	ô	244	ő	525	ō	333
ú	250	ù	249	û	251	ű	533	ū	363
ŕ	341	ṙ	7769	r̂	531	r̋	529	ṛ	7771

Abb. 36: Die Kodierungstabelle

Als Schriftart für die Darstellung der Akzentsymbole dient DoulosSil IPA, der jederzeit aus Gralis heruntergeladen werden kann. Die Schreibung der Akzentzeichen kann auf zwei Arten erfolgen: Eine liegt darin, dass mittels Klick das Akzentzeichen oder die posttonale Länge für jeden Vokal einzeln aufgerufen wird, wodurch im Unicode-Standard alle vier Akzente (lang steigend, kurz steigend, lang fallend, kurz fallend) und die Länge dargestellt werden kann.

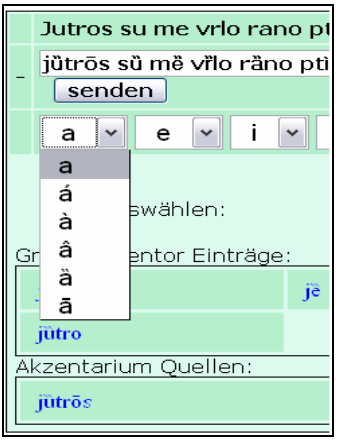


Abb. 37: Einfügen eines Akzentes per Klick

Die zweite Möglichkeit zum Einfügen der Akzente besteht darin, nach dem Akzent eines entsprechenden Wortes zu suchen, wie im hier dargestellten Fall jenen des Lexems *jutro*.

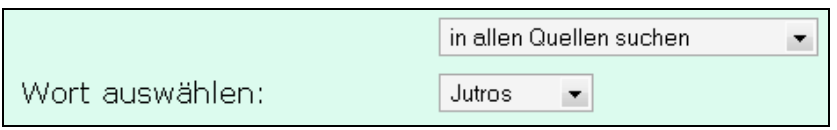


Abb. 38: Einfügen eines Akzentes durch Eintragen eines Wortes

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

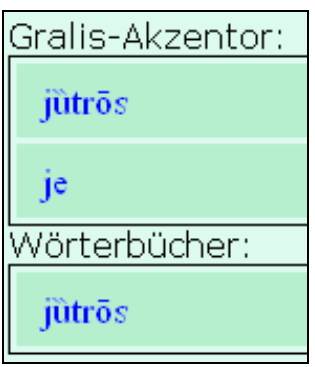


Abb. 39: Anzeige von Lexemen aus dem Speech-Korpus

Die nun in ihren Einzelheiten beschriebene Maske des Gralis-Akzentariums bestitzt in ihrer Gesamtheit folgendes Aussehen:

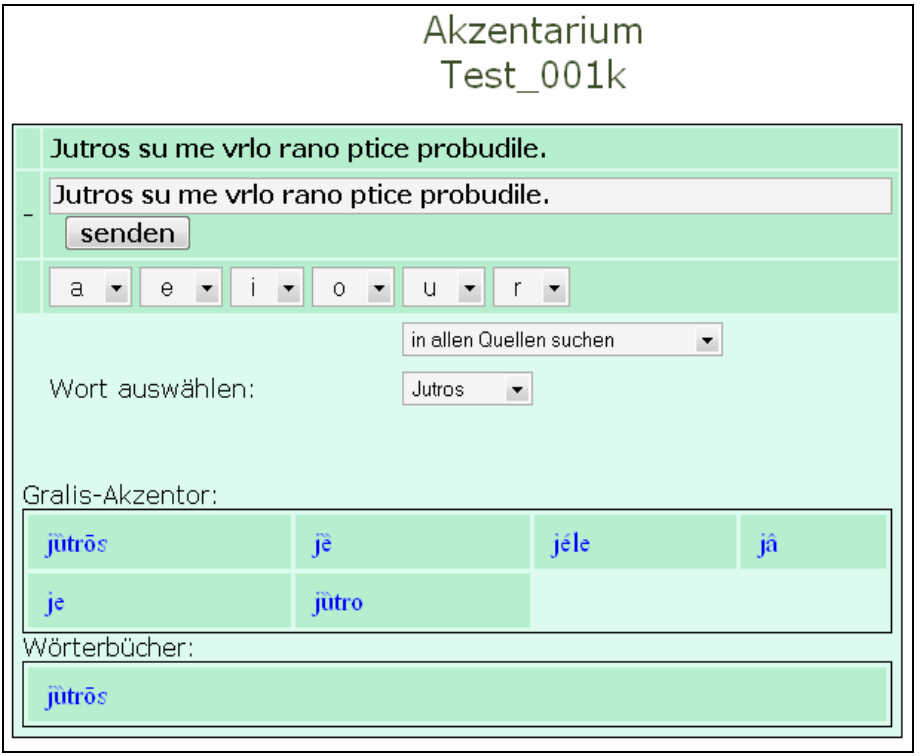


Abb. 40: Gesamtansicht des Gralis-Akzentariums

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.

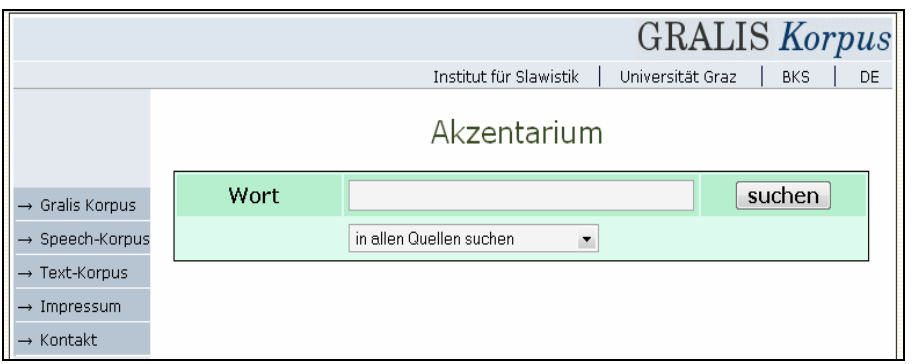


Abb. 41: Das Suchinterface des Gralis-Akzentariums

Ein in dieses Fenster eingetragenes Wort (bzw. auch eine Wortform) kann sodann in sämtlichen Quellen des Gralis-Akzentariums gesucht werden, wobei bei Anwahl dieser Option und Einfügen eines nicht akzentuierten Wortes (hier: *televizija*) folgendes Ergebnis zu Tage tritt:

Abb. 42: Die Suche in ausgewählten Quellen

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).

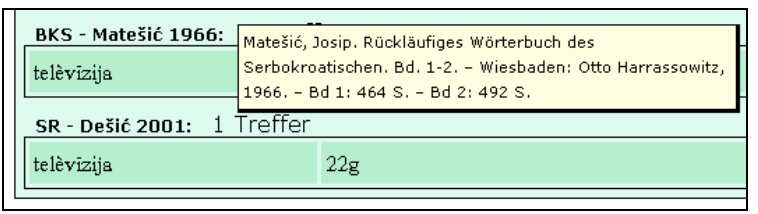


Abb. 43: Beispiel einer Quellenangabe

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).

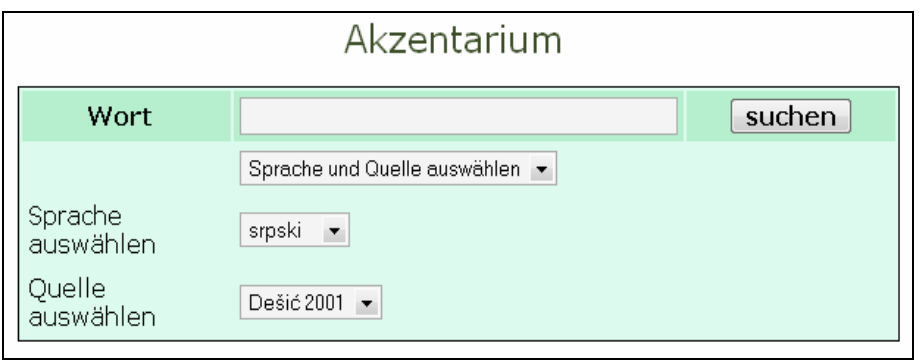


Abb. 44: Wahl einer bestimmten Quelle

Die Aufbereitung von ins Gralis-Akzentarium einzufügenden Lexemen erfolgt im Programm Word, wozu zuallererst eine Wortliste als txt-Datei abgespeichert wird, woraufhin die Akzentuierung mittels Betätigen der Tastenkombinationen Alt + 1, 2, 3, 4 oder 5 vorgenommen werden kann.

Kombinacija		Akcenat	Znak	Zeichen-Code: Unicode (hex)	Primjer
Alt	1	dugouzlazni	◌́	0301	rúka
	2	kratkouzlazni	◌̀	0300	nòga
	3	dugosilazni	◌̂	0302	maĵka
	4	kratkosilazni	◌̏	030F	saȍn
	5	postakcenatska dužina	◌̄	0304	žénā

Abb. 45: Darstellung der Tastenkombinationen zur Niederschrift der Akzente

Angesichts dessen, dass die Akzente auf dem benutzten Server im Unicode-Zeichensatz und nicht als Kombination zweier Zeichen dargestellt werden, muss in allen Fällen eine Dekodierung der Kombinationen vorgenommen werden, wozu ein Makro namens Akzent-Unicode geschaffen wurde, das zwei (durch die Kombination Alt + 1, 2, 3, 4 oder 5) erhaltene Zeichen automatisch in ein Unicode-Zeichen umwandelt.

Sollte eine von TrägerInnen der Urheberrechte zur Verfügung gestellte Wortliste nicht gemäß dem Unicode-Standard erstellt worden sein, kann mit dem bereits erwähnten Makro Akzent-Unicode eine Dekodierung in das erforderliche Format durchgeführt werden. Der nächste Schritt liegt darin, die gesamte Wortliste in eine einspaltige Tabelle einzufügen, die sodann in eine bestehende Tabelle mit zwei Spalten übertragen wird. Der Inhalt der zweiten Spalte wird daraufhin verborgen, sodass einzig die erste Spalte sichtbar bleibt, in der die Akzentzeichen der akzentuierten Lexeme mithilfe des genannten Makros entfernt werden. Erst zu diesem Zeitpunkt kann die verborgene Spalte erneut aktiviert und angezeigt werden. Auf diese Weise erhält man zwei Spalten, wobei die erste nicht akzentuierte und die zweite akzentuierte Wörter und Wortformen enthält. Diese hier beschriebenen Arbeitsschritte sind deshalb erforderlich, damit im Gralis-Akzentarium eine Suche nach akzentuierten Lexemen durch die Eingabe von nicht akzentuierten Formen möglich wird.

Für die Zukunft ist eine Ausweitung des Gralis-Akzentariums auf die beiden anderen Studienrichtungssprachen des Grazer Institutes für Slawistik geplant (russisch und slowenisch), die mit dem Akzentarium für die Sprachen bosnisch/bosniakisch, kroatisch und serbisch in direktem Zusammenhang stehen werden, wodurch ein Medium für effiziente und technisch leicht durchführbare kontrastive Analysen geschaffen werden soll.

Olga Lehner (Graz)

Die technische Entwicklung des Gralis Speech-Korpus

15. In der folgenden Darstellung werden die technischen Aspekte der Entwicklung des Gralis Speech-Korpus erläutert, der der Sammlung, Auswertung und Annotation von Audioaufnahmen dient. Die Korpora werden in einer MySQL-Datenbank verwaltet, die neben den Daten auch den gesamten Entwicklungsprozess abbildet. Die Benutzer- und Administrationsschnittstellen werden, nach Benutzerklassen aufgegliedert, jeweils über ein PHP-Webinterface realisiert.

L a m p . Das Gralis Speech-Korpus wurde auf einem Apache-Web-Server unter dem Betriebssystem Linux generiert, wobei die Information in einer MySQL-Datenbank gespeichert und über ein Webinterface administriert wird. Diese Implikation wie auch die Erstellung von dynamischen Inhalten wurde in PHP programmiert. PHP (rekursives Akronym für „Hypertext Preprocessor“, ursprünglich „Personal Home Page Tools“) ist eine serverseitig interpretierte Skriptsprache mit einer an C bzw. C++ angelehnten Syntax mit breiter Datenbankunterstützung und Internet-Protokolleinbindung. Häufig wird diese Open-Source Kombination (Linux-Apache-MySQL-PHP) als Lamp bezeichnet.

U T F - 8 - K o d i e r u n g . Sowohl auf den HTML-Seiten als auch in der MySQL-Datenbank wird die Unicode UTF-8-Kodierung verwendet (Abk. für das 8-bit Unicode Transformation Format), die bis zu vier Byte unterstützt und auf die sich wie bei allen UTF-Formaten sämtliche 1.114.112 Unicode-Zeichen abbilden lassen. UTF-8 beinhaltet auch die IPA-Symbole (International Phonetic Alphabet), die wir für Akzente, Transkriptionen und Intonationen verwenden. Für die Benutzung reicht es aus, die frei zugängliche Unicodeschrift DoulosSil (http://scripts.sil.org/DoulosSIL_download) zu installieren.

A r b e i t s u m g e b u n g . Die gewählten Arbeitsinstrumente ermöglichen uns, ein eigenes Administrationssystem und eine bequeme Arbeitsumgebung für die am Projekt mitarbeitenden Personen zu schaffen und auf eventuell auftretende Schwierigkeiten, neue Bedürfnisse und Ideen flexibel reagieren zu können.

Die Benutzer des Programms sind auf Gruppen mit verschiedenen Aufgaben und Zugriffsrechten aufgeteilt: Administrator, Mitarbeiter/In, Studierende und ExpertInnen.

Jede(r) Mitarbeiter/In hat eine eigene Arbeitsschnittstelle, wo ihr/ihm die volle Information über die Mitarbeit am Korpus gewährt wird: Anzahl und Liste der eingetragenen Aufnahmeevidenzen, Information über hochgeladene, vollständige und segmentierte wav- und mp3-Audiodateien, über durchgeführte Annotationen der Aufnahmen (Akzente, Transkription und Intonation).

Zu jeder Zeit haben die Mitarbeitenden die Möglichkeit, Aufnahmeevidenzen zu korrigieren, zu löschen, ein unvollendetes Formular fertig zu stellen oder eine neue Aufzeichnung für eine bereits existierende Aufnahme zu ergänzen. Zur Beschleunigung der Eingabe kann man schon existierende Evidenzblätter als Vorlage für den neuen Eintrag heranziehen, was insbesondere bei der Bearbeitung von Sprechenden mit ähnlichen Biographien nützlich ist, wie zum Beispiel bei Schülern einer Klasse, Bewohner eines Ortes etc.

Aus der Gesamtliste können Aufnahmen ausgewählt werden, die sich entweder nur auf das Wort-Korpus (Liste isoliert ausgesprochener Wörter) oder einzig auf das Fix-Korpus beziehen. Die Liste der bearbeiteten Aufnahmeevidenzen kann man nach der laufenden Nummer der Aufzeichnung oder des Sprecher, der Chiffre und dem Datum der Eingabe des Formulars sortieren.

Für die Verwaltung des Korpus erhält der Administrator umfassende Informationen über die bearbeiteten Aufnahmeevidenzen sowie über den jeweiligen Arbeitsfortgang der mitarbeitenden Personen.

p h p M y A d m i n . Als zusätzliches Werkzeug für die Administration der MySQL-Datenbank verfügt der Administrator über das Programm phpMyAdmin (http://www.phpmyadmin.net), das unter der GNU General Public License lizenziert ist. Dieses Programm stellt ein PHP-basiertes MySQL-Administrierungssystem dar, d. h. es bietet die Möglichkeit, über einen gewöhnlichen Web-Browser eine MySQL-Datenbank zu verwalten (Tabellen und Datensätze anlegen, ändern und löschen, SQL-Befehle ausführen, Datenbankdateien in verschiedenen Formaten exportieren).

D a t e n b a n k s t r u k t u r . Die Datenbank enthält einige allgemeine Tabellen mit Angaben zur Aufnahmeevidenz und Charakteristik der aufgenommenen Personen: geographische Bezeichnungen (Länder, Regionen, Orte), Muttersprache(n) und Fremdsprachen, Idiome, Mundarten, Nationalität, religiöse Zugehörigkeit, Beruf usw.

Weitere Tabellen enthalten Informationen über die Beschreibung der Audioaufnahme wie Analyse, Typ und Marke des Apparats, der für die Abnahme der Aufzeichnung verwendet wurde, weiters unter welchen Bedingungen die Aufzeichnung stattfand u. Ä. Bei Bedarf fügen die am Korpus mitarbeitenden Personen mit Hilfe des Webinterfaces zu diesen Tabellen neue Angaben ein, die sofort im Aufnahmeevidenzformular zugänglich werden.

Es folgen Tabellen mit Angaben über die in allen Fällen stets anonymen aufgenommenen Personen, eine Tabelle mit der Information zu den Audiodateien der vollen (gesamten) und (in Sätze) segmentierten Aufnahmen.

Das Korpus wird durch mehrere Tabellen dargestellt: Die Thementabelle (*Jutro*, *Na moru*, *Odlazak na ispit*, *Moji roditelji* usw.), die Satz- und Wörtertabelle, die Tabelle der Wörter mit kanonischem (erwartetem) Akzent,

Gralis-Akzentor – mit dem tatsächliche gesprochenen Akzent, Gralis-Transkriptor – Tabelle der transkribierten Sätze und Wörter, Supersegmentarium – Tabelle der Sätze mit Intonation. Durch die Wahl einer solchen Datenbankstruktur erhält man wesentlich umfassendere Möglichkeiten zur Durchführung von Suchabfragen.

S u c h a b f r a g e n . Wenn man zum Beispiel im Rahmen der einfachen Suche innerhalb des Fix-Korpus einen der Sätze ausgewählt hat, wird die Auswahl nach fünf Kriterien durchgeführt, die einen Sprechenden grundlegend charakterisieren (Geschlecht, Nationalität, Geburtsort und Lebensmittelpunkte, Muttersprache).

In der erweiterten Suche kann man sodann eine Auswahl zu insgesamt über 25 Parametern durchführen. Als Ergebnis der Suche wird die Möglichkeit geboten, den gewählten Satz von verschiedenen Sprechenden anzuhören, die die eingegebenen Kriterien erfüllen (z. B.: weibliche Sprechende mit XXX-Muttersprache).

Durch Klicken auf das Symbol „Tabelle“ in jeder Zeile der Ergebnisstabelle öffnet sich ein neues Fenster, in dem die volle Liste der Sätze inklusive Links auf die entsprechenden Audiodateien, die zum selben Thema gehören und von der gleichen Person gelesen wurden, angezeigt wird.

Man kann die Suche auf annotierte Aufnahmen begrenzen, das heißt auf diejenigen Aufzeichnungen, in denen für jedes Wort der Akzent und die Transkription vermerkt sowie die Intonation markiert wurden. Angesichts der Tatsache, dass die Annotation einen überaus arbeitsintensiven Prozess darstellt, ist es nur schwer möglich, das gesamte Korpus zu annotieren.

Das ebenfalls im Rahmen des Gralis Speech-Korpus entwickelte Wort-Korpus besteht aus Aufnahmen mit isoliert ausgesprochenen Wörtern, die in Wortlisten zusammengefasst wurden. In Ergänzung zur Auswahl nach dem Sprechenden kann eine Suche nach einem Wort, z. B. *balon*, durchgeführt werden, wobei additional auch eine Eingabe mit Wildcard, z. B. **ba***, möglich ist.

Im Sinne der Erzielung einer größtmöglichen Einheitlichkeit innerhalb des Gralis-Speech-Korpus ist es in absehbarer Zukunft geplant, für das Korpus dieselbe Suchsyntax wie in der IMS Corpus Workbench, auf der das Gralis Text-Korpus fußt, zur Anwendung zu bringen.

Alexander Just – Arno Wonisch (Graz)

Aufnahme, Dekodierung, Bearbeitung und Upload von Audiodateien in das Gralis Fix-Korpus

16. Im Zuge einer Aufnahme, Dekodierung in ein gewünschtes Format, Bearbeitung bzw. Optimierung und schließlich Auswertung bzw. Analyse von Audiodateien gilt es mehrere Arbeitsschritte in unterschiedlichen Programmen durchzuführen, die in vorliegender Anleitung in chronologischer Abfolge dargestellt werden.

Aufnahme von Audiodateien. Zur Durchführung dieses ersten Arbeitsschrittes stehen insgesamt zwei Diktiergeräte und ein Audiorecorder zur Verfügung, die Aufnahmen in folgenden Formaten ermöglichen:

(a) Zwei Diktiergeräte des Typs Olympus WS-100 – Format: wma (eignet sich für Audioanalysen)

(b) Audiorecorder Edirol R-09 – Format: wav (eignet sich für Audio- aber auch für Spektralanalysen)

(c) Weiters bedienen wir uns des Programms Skype Recorder, mit dem via Skype geführte Internettelefonate aufgezeichnet werden können, sodass eine Vielzahl der Aufnahmen direkt vom Arbeitsplatz aus getätigt werden.

Dekodierung vom Format wma ins Format mp3 mithilfe des MediaCoders. Der „MediaCoder" dient ausschließlich dazu, Audiofiles aus einem Format (wma, wav, mp3, ogg) in ein anderes umzuwandeln. Nach der Transformation der gewünschten Audiodatei ins Format wav kann mit der Bearbeitung bzw. Analyse begonnen werden, die in einem ersten Arbeitsschritt im Programm WaveLab vorgenommen wird.

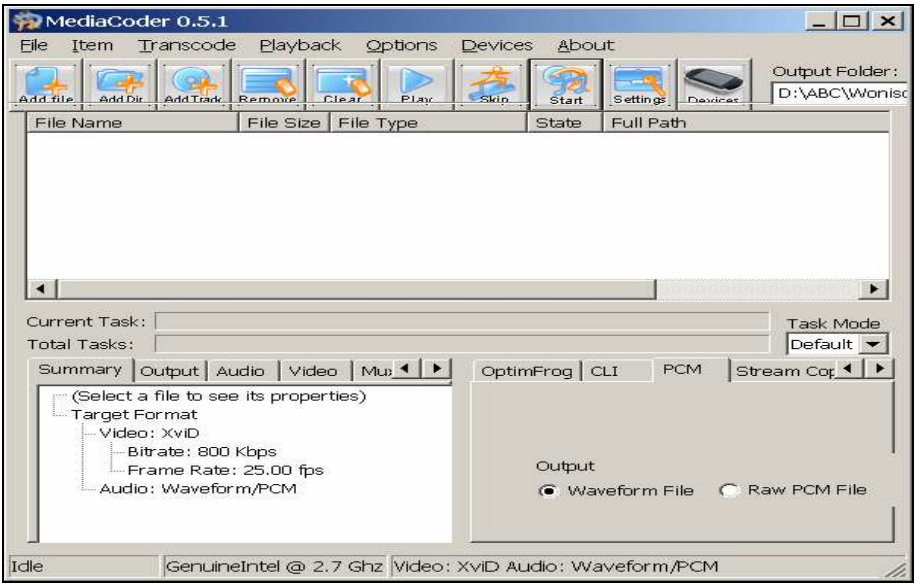


Abb. 46: Arbeitsoberfläche des Programms „MediaCoder“

Bearbeitung und Segmentieren von Audiodateien: Das Programm WaveLab. Das von „Steinberg Media Technologies GmbH“ (www.steinberg.de) entwickelte Programm WaveLab dient im Rahmen des Gralis-Speech-Korpus (bestehend aus dem Gralis-Wort-Korpus, dem Gralis-Fix-Korpus und dem Gralis-Frei-Korpus) zur Be- und Verarbeitung der aufgenommenen Audiodateien. Bei der im Rahmen des Projektes „Die Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen“ verwendeten Version handelt es sich um WaveLab 6, das vom Institut für Slawistik der Karl-Franzens-Universität Graz im Jahre 2007 erworben wurde.

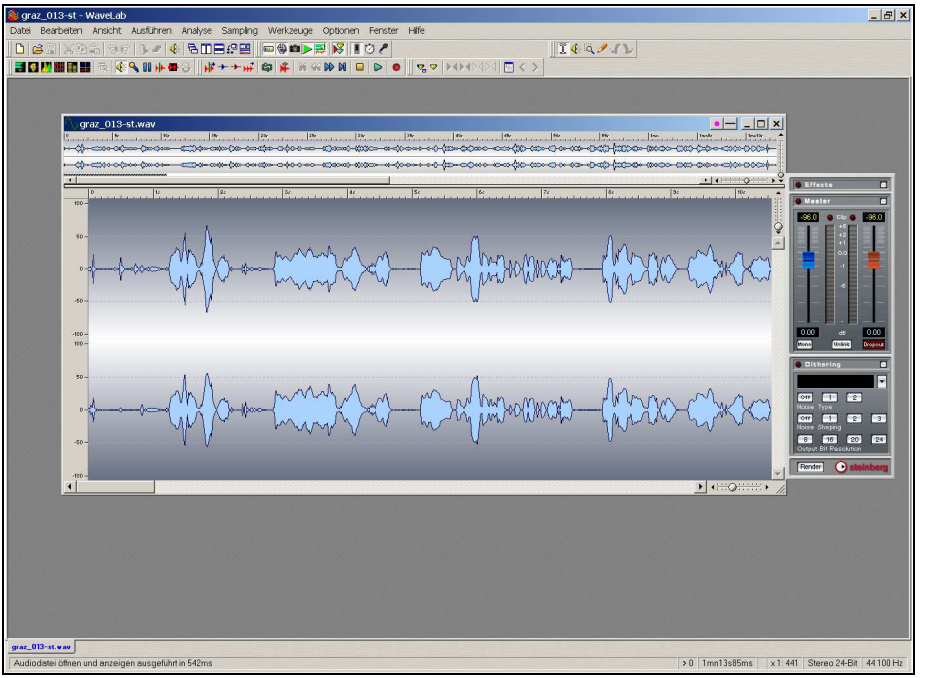


Abb. 47: Darstellung von Stereotonspuren im Programm „WaveLab“

Segmentieren bzw. Splitten von Audiodateien im Programm WaveLab. Das Splitten erfolgt mittels Markern, wobei der von uns gewählte Marker der gelbe Standardmarker in der dritten Menüleiste von oben ist. In einem ersten Arbeitsschritt werden die zu segmentierenden Einheiten (Sätze, Wörter) durch ein Setzen des „Standardmarkers“ an den entsprechenden Stellen markiert. Sodann ist der in Abb. 3 dargestellte Befehl „Wave in aktivem Fenster“ anzuwählen, der sich auf die gerade geöffnete Arbeitsoberfläche bezieht.

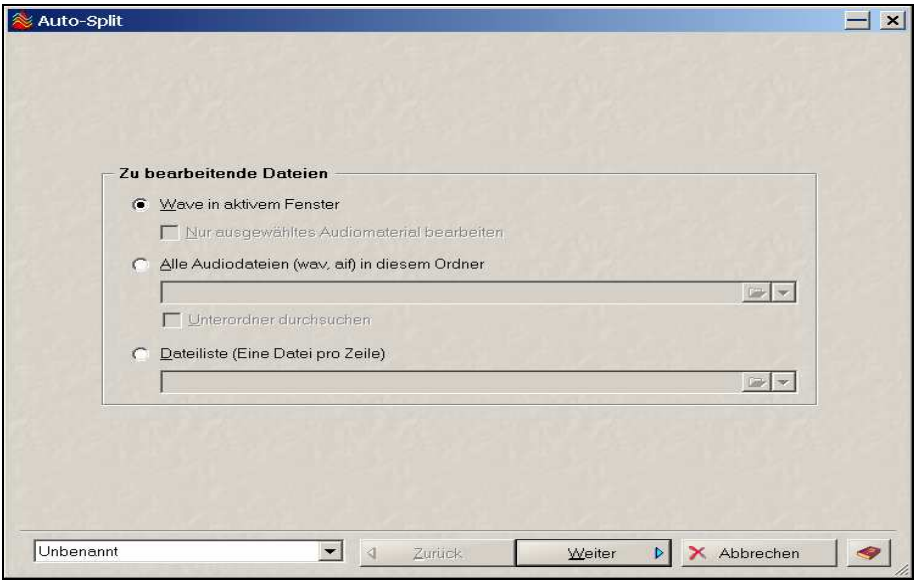


Abb. 48: Auswahl „Wave in aktivem Fenster“ im geöffneten Arbeitsfenster

In weiterer Folge ist im nächsten Fenster die Option „Auto-Split entsprechend der Marker“ [sic!] zu aktivieren, damit die Segmentierung gemäß den gesetzten Markern erfolgen kann.

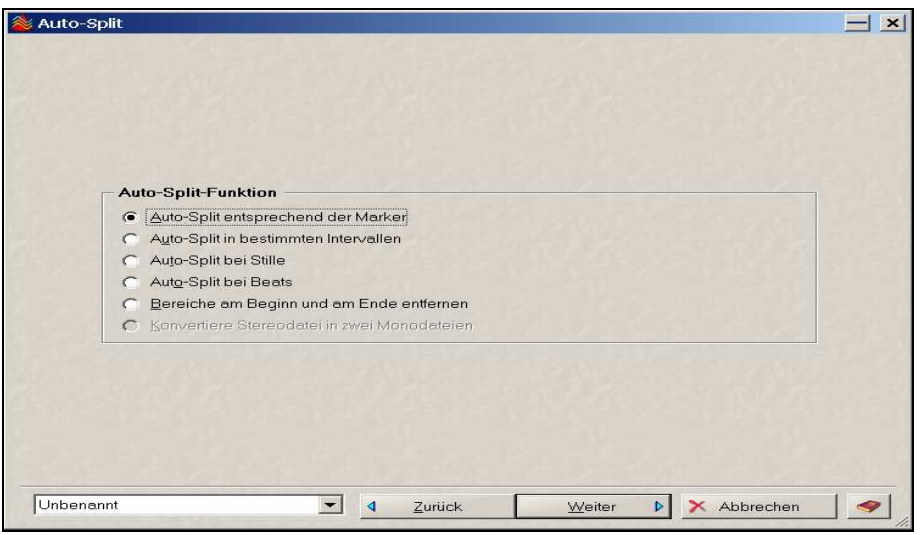


Abb. 49: Segmentierung gemäß den Markern

Im Zuge der Bearbeitung der in das Gralis Speech-Korpus hochzuladenden Dateien erwies es sich angesichts der Struktur der Audiofiles als zweckmäßig, zur Segmentierung die vom Programm primär angebotenen Standard-Marker heranzuziehen.

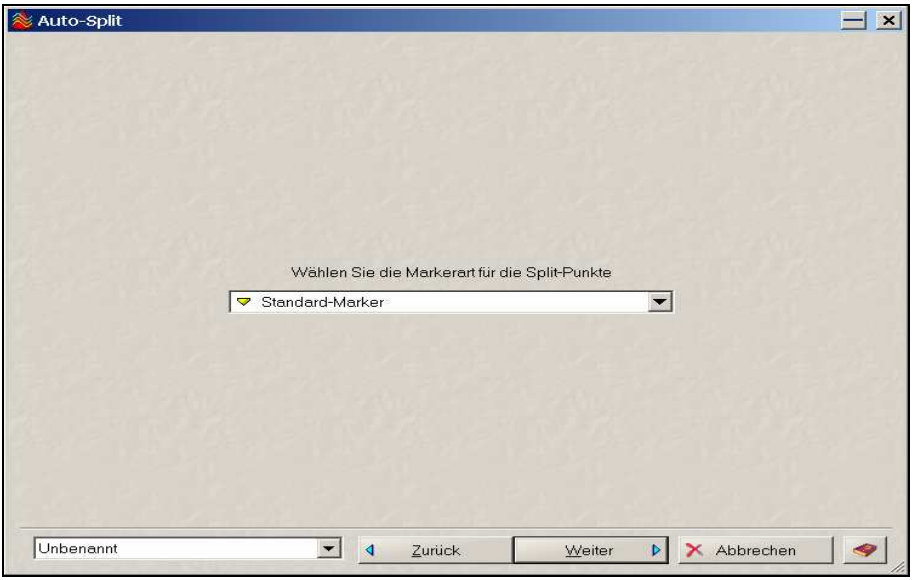


Abb. 50: Auto-Split entsprechend den gesetzten (Standard)-Markern

Wichtig ist in weiterer Folge die Wahl des entsprechenden „Zielordners“, in dem die Datei abgespeichert werden soll, wozu in der oben stehenden weißen Zeile dieser auszuwählen ist. Die übrigen Optionen auf dieser Seite sind freizulassen bzw. entsprechend den Voreinstellungen zu übernehmen.

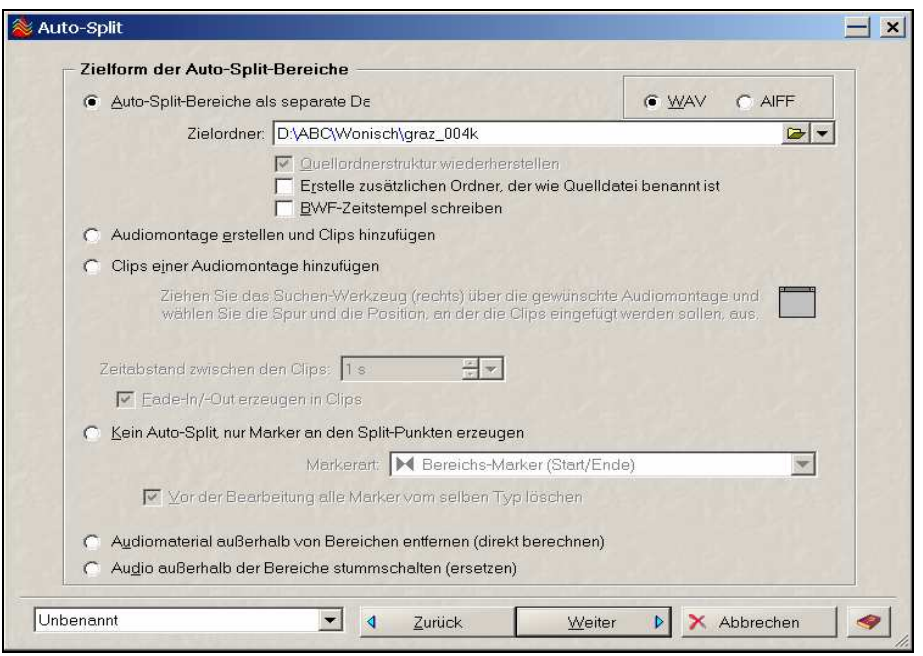


Abb. 51: Wahl des Zielordners

Durch einen Klick auf „Weiter“ gelangt man zu folgendem Fenster in dem sich die Wahl einer Stille von 0 s 500 ms zu Beginn und am Ende einer Datei empfiehlt, wodurch ein Abhören und eine Analyse einer Audiodatei mit einer zeitlichen Verzögerung von einer halben Sekunde vorgenommen werden kann.

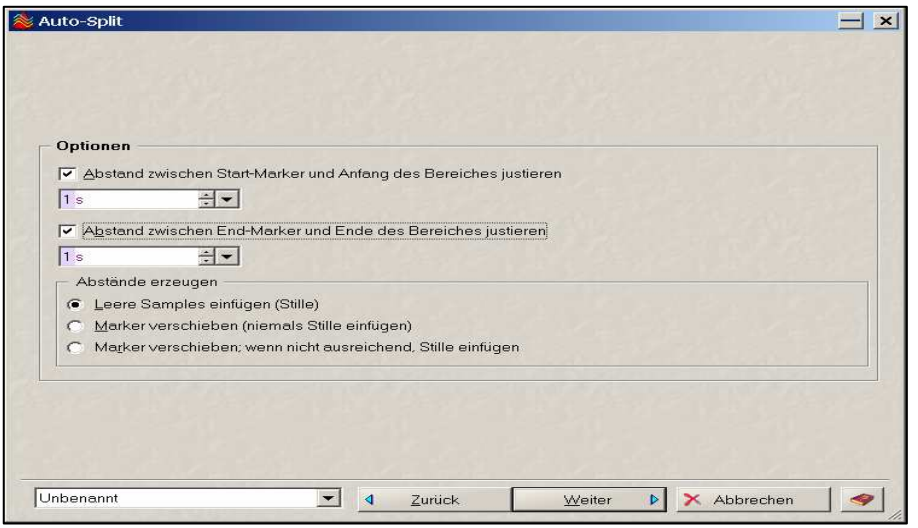


Abb. 52: Einfügen von Stille zu Beginn und am Ende der segmentierten Audiodatei

Danach sind in ein Fenster die zu den Audiodateien (d. h. Sätzen) gehörenden Benennungen gemäß dem vereinbarten Schema für Chiffren einzufügen, wobei eine Zeile hier für einen Satz steht und der Befehl „Liste (ein Name pro Zeile)“ zu wählen ist.

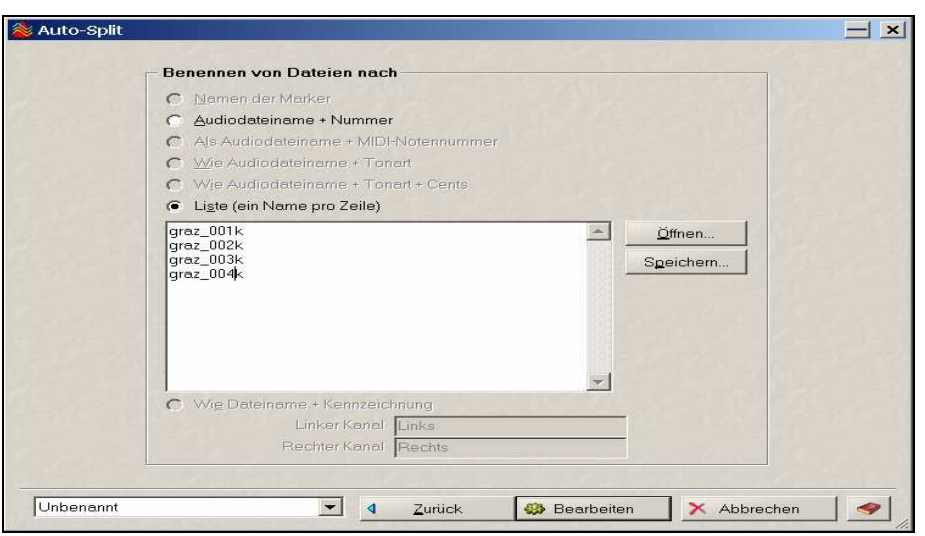


Abb. 53: Wahl der Dateinamen

Durch einen abschließenden Klick auf „Bearbeiten“ wird die Segmentierung durchgeführt, woraufhin mit einem abschließenden Speichern sind alle Arbeitsschritte abgeschlossen sind und die segmentierten Einheiten unter den in Abb. 8 gewählten Dateinamen abgespeichert wurden.

Analyse von Audiodateien: Das Programm Praat. Das am Institute of Phonetic Sciences an der Universität Amsterdam von Paul Boersma und David Weenink entwickelte Open-Source-Programm Programm Praat (dt. Übersetzung: „sprechen“) dient zur akustischen Analyse von Audiomaterial im Format wav.

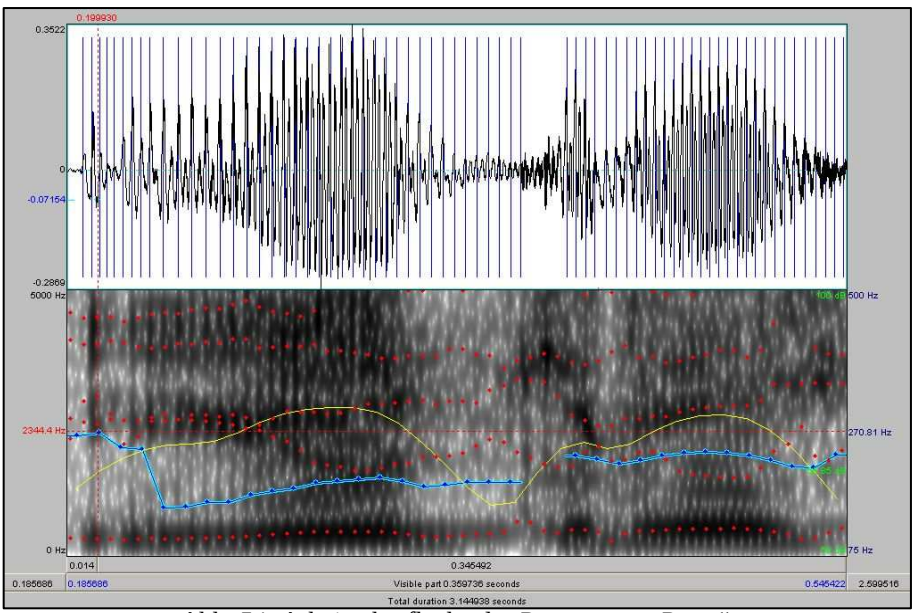


Abb. 54: Arbeitsoberfläche des Programms „Praat“

Praat (http://www.fon.hum.uva.nl/praat/) 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.

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, wurde unter der Leitung von Piet Martens entwickelt und 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 Annotation 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.

Transkription von Audiodateien: Das Programm Adaba. Dieses Programm stellt das Ergebnis eines sechsjährigen Forschungsprojektes dar und wurde dankenswerterweise vom Leiter dieses Projektes, Rudolf Muhr vom Institut für Germanistik der Universität Graz, zur Verfügung gestellt. Es findet Verwendung bei der Transkription gesprochener, isolierter Wörter, d. h. im Zuge der Bearbeitungsschritte im Rahmen des Gralis Wort-Korpus. Ursprünglich wurde Adaba für einen phonetischen Vergleich von Wörtern entwickelt, die von jeweils einer weiblichen Sprecherin und einem männlichen Sprecher aus Österreich, Deutschland und der Schweiz artikuliert werden. Neben der Audioimplikation verfügt Adaba über eine gemäß dem internationalen phonetischen Alphabet IPA erstellte Transkription sämtlicher Wörter.

Boris Tošović (Graz)

Die Gralis Audio-VideoTools

17. 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.

Das Gralis Sat-Tools-Skript wurde im Rahmen des Projekts „Die Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen“ entwickelt und dient unter anderem zur automatischen Steuerung und Bedienung des Programms ProjectX, wobei alle Programme über eine graphische Benutzeroberfläche (GUI) bedient werden können.

Abb. 55: Die graphische Benutzeroberfläche von Gralis Sat-Tools-Skript RC1

Das Skript kann dabei mit folgenden Programmen benützt werden: MPEG2-Schnitt, Cutterman, MuxMan, IMAGO-MPEG muxer, notepad++, SubtitleCreator und IfoEdit. In ProjectX ist es möglich andere Programme und Befehle auszuführen, Postprocessing, wodurch das Skript flexibel eingesetzt und erweitert werden kann.

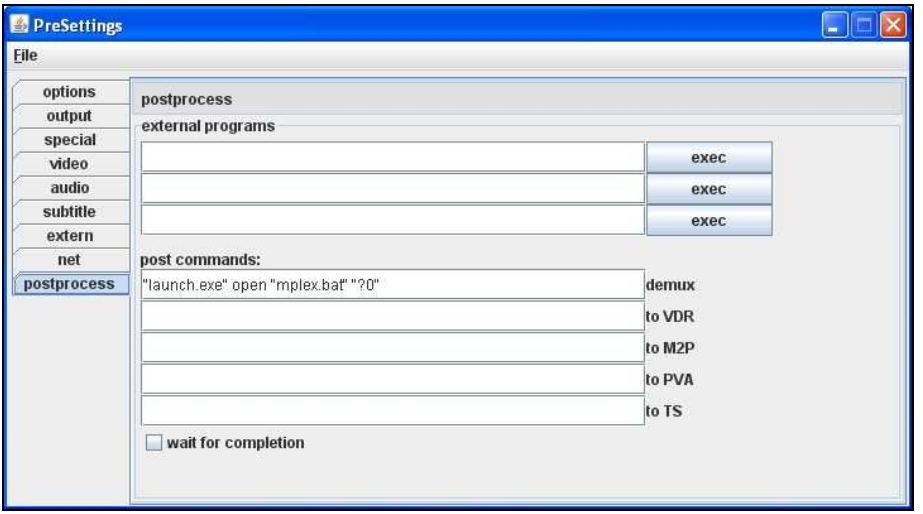


Abb. 56: Einstellungen in ProjectX, Postprocessing

Das Skript der Gralis Audio-VideoTools wurde an die Erfordernisse des genannten Projektes angepasst. Ein Beispiel dafür ist die automatische Bearbeitung mit ProjectX und mplex, wobei zuerst ProjectX die Bearbeitung durchführt und mittels mplex eine mpeg-Datei oder eine DVD-Struktur erzeugt werden kann. Dieser Arbeitsschritt stellt sich im Skript wie folgt dar:

```
; PostCommands
$post_cmd1 = 'PostCommands.Cmd4="' & @ScriptDir & '\' & '$launch_exe"' & ' open "' & @ScriptDir & '\' & '$mplex_bat"' & ' "?0"'
$post_cmd2 = 'PostCommands.Cmd5="' & @ScriptDir & '\' & '$launch_exe"' & ' open "' & @ScriptDir & '\' & '$batchmux_cmd"' & ' "?0"'
```

Abb. 57: Skriptdetail, Ausführung von mplex

In der Benutzeroberfläche können Variablen leicht verändert werden, wodurch auch der Grad der Automatisierung von Prozessen ausgewählt werden kann.

Die DVB-S-Streams wurden im RAW-Format aufgenommen – ein Format, in dem alle Sat-TV Kanäle (ausgenommen HDTV) auf DVB-S in MPEG-2 gesendet werden. Zur Komprimierungen des Videos werden freie Codecs wie XVID und x264 benutzt. MPEG2 stellt einen generischen MPEG-Standard zur Videodekodierung mit Videokompression wie auch zur Audiokodierung mit Audiokompression dar. Die Einführung von MPEG-2 erfolgte im Jahre 1994 als Weiterentwicklung von MPEG-1. Generell bedeutet dies, dass ein Datenformat und ein Dekodierungsverfahren festgelegt werden, ohne dabei andere Parameter wie z. B. die Auflösung zu bestimmen. DVB-S, DVB-C und DVB-T benutzen allesamt das Format MPEG-2.

XVID ist ein Open-Source-MPEG-4-Video-Codec, der ursprünglich auf dem OpenDivX-Quelltext basierte, wobei als Quellcode OpenDivX als

Ergebnis der MPEG-4-Referenzimplementierung des EU-Projekts MoMuSys herangezogen wurde. Nach der Schließung des Quellcodes von OpenDivX wurde das XviD-Projekt von mehreren freiwilligen ProgrammentwicklerInnen gestartet, die durch den unverschlüsselt veröffentlichten Quelltext von OpenDivX die Möglichkeit erhielten, den Codec in seinen grundlegenden Eigenschaften zu verändern und damit zu optimieren. Die bekanntesten MPEG-4-Encoder sind XviD, DivX und Nero Digital.

x264 ist ein freies und plattformübergreifendes Kodierungsprogramm für das Video-Format H.264 (MPEG-4 AVC) und wird unter der GNU (General Public License) zur Verfügung gestellt. Das Programm x264 befindet sich derzeit noch in der Entwicklungsphase, die beinahe täglich neue Versionen hervorbringt, weshalb es als Software im Alpha-Stadium anzusehen ist. Die x264-Kodierer-Bibliothek mit einem Programmcode in ISO-C wird von Laurent Aimar, Eric Petit, Loren Merritt, Min Chen, Måns Rullgård, Justin Clay, Radek Czyz, Alex Izvorski, Christian Heine und Alex Wright bearbeitet und weiterentwickelt. Als H.264-Kodierer erweist sich x264 als deutlich effizienter als diejenigen Codecs, die auf dem einfacheren MPEG-4 ASP basieren. Der x264-Codec beinhaltet keinen Decoder wie z. B. DivX oder XviD, und um mit x264 komprimierte Videos abspielen zu können, benötigt man zusätzliche Software.

Benutzte Programme: P r o j e c t X . Beim digitalen Fernsehen wird das Fernsehbild vor der Ausstrahlung komprimiert und als digitaler Datenstrom (MPEG2 Transportstream) gesendet. Der MPEG2 Transportstream gehört zwar zur Familie der MPEG2-Standards, sollte jedoch vor dem Weiterverarbeiten konvertiert werden, da die Struktur des Aufbaus eines MPEG2 Transportstreams eine andere ist.

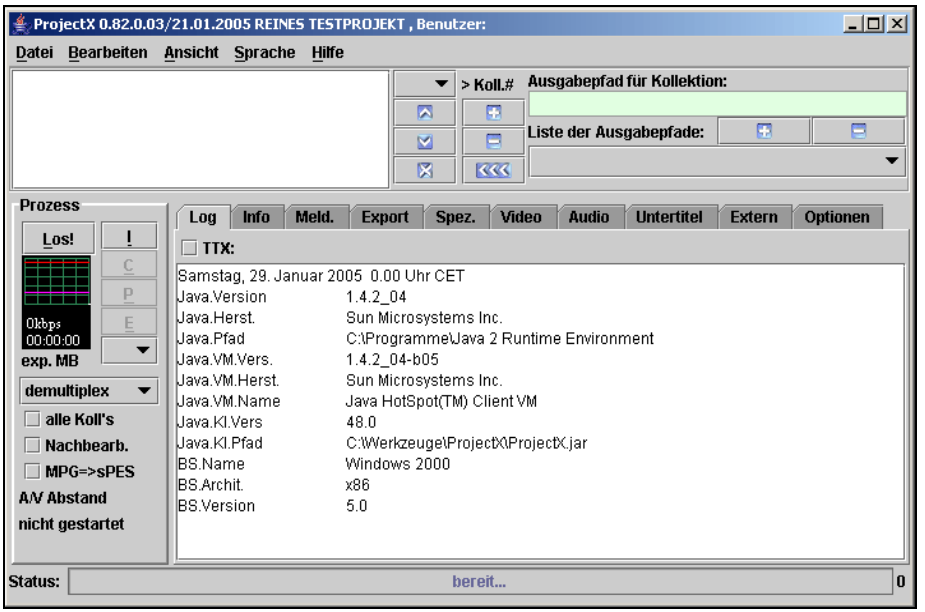


Abb. 58: Das graphische Interface von ProjectX

An dieser Stelle kommt das Programm ProjectX mit seiner oben dargestellten Benutzeroberfläche ins Spiel. Die Arbeitsweise in diesem Programm ähnelt jener in einem MPEG2-TS-Hardware-Decoder, indem Project X den MPEG2-Transport-Stream in seine Bestandteile (MPEG Elementary Streams, Video-, Audio- und sonstige Daten) zerlegt. Beim Trennen wird der MPEG2-Transport-Stream auf Fehler im Stream überprüft und synchronisiert, während beim Konvertieren nur eine grobe Prüfung stattfindet. Folgende Transportstreamformate (sowohl einfache als auch multiple TS und PS) werden vom Programm unterstützt.

DVB MPEG2 Transportstream (DVB MPEG2-TS), MPTS (Multiple Program Transport Stream),

Packet Video Audio (PVA, PSV, PSA, PAV),

MPEG Program Stream (MPEG1/2 PS),

Linux Video Disc Recorder (Linux VDR),

Packetized Elementary Stream (PES RAW Streams),

Elementary Stream (ES Streams).

ProjectX ist ein freies Bildungs- und Testprojekt, das auf der Programmiersprache Java basiert. Das Programm wird in Form von Open Source Codes verteilt und muss von den BenutzerInnen bei Bedarf selbst kompiliert werden, wobei es sowohl über ein GUI (Graphical User Interface) als auch CLI (Command Line Interface) bedient werden kann. Im Folgenden seien die

wichtigsten Programme genannt, die im Rahmen von ProjektX zur Anwendung kommen.

M P E G 2 S c h n i t t u n d C u t t e r m a n . Diese zwei Programme werden für die Segmentierung bzw. das Schneiden der Videodateien verwendet. m e G U I stellt eine graphische Benutzeroberfläche dar, die unterschiedlichste Programme zur Videobearbeitung und Audio Bearbeitung zusammenfasst und somit eine komplett automatisierte Video- und Audio-Bearbeitung ermöglicht. Der M e d i a C o d e r schließlich ist ein freier Medientranscoder, der für die gängigsten Audio- und Video-Codecs und -Tools benutzt werden kann.

A u d i o f o r m a t e i m G r a l i s S p e e c h - K o r p u s . Bei einem Audioformat handelt es sich um ein Dateiformat, das den Aufbau einer Audiodatei und dabei meistens komplette Klangkurven eines Tonsignals beschreibt. Man unterscheidet zwei Arten von komprimierten Daten, nämlich verlustfreie und verlustbehaftete, wobei unkomprimierte, verlustfreie sowie verlustbehaftete, komprimierte Formate in verschiedenen Containerdateien untergebracht werden können, wie etwa: MP4, Matroska, Ogg oder WAV. Unkomprimierte Formate sind z. B. Rohdaten (pcm), Audio Interchange File Format (AIFF, Containerformat) und Interchange File Format: RIFF WAVE (Containerformat) (wav). Zu den verlustfreien komprimierenden Verfahren gehören etwa MPEG-4 Audio Lossless Coding (MPEG-4 ALS), Apple Lossless Audio Codec (ALAC), AudioZip, Free Lossless Audio Codec (flac, fla, ggf und ogg), RealAudio Lossless (ra, in Matroska-Containern ggf und mka) und Windows Media Audio Lossless (wma). Verlustbehaftet komprimierende Verfahren sind z. B. Dolby Digital, A/52 oder AC3 (ac3), DTS Coherent Acoustics (dts), Adaptive Transform Acoustic Coding (ATRAC), MPEG 1/2/2.5 Audio (mp1, mp2, mp3), MPEG 2/4 Audio (aac, mp4, m4a), RealAudio (ra), Windows Media Audio (wma) und Vorbis (ogg). Zu den Sprachcodecs zählen unter anderem Speex (spx, ggf und ogg) oder Digital Speech Standard (DSS).

Im Folgenden soll versucht werden, die wichtigsten Formate für das Gralis-Korpus darzustellen. Bei R I F F W A V E / P u l s C o d e M o d u l a t i o n (P C M) handelt es sich um ein WAVE-Format, das auf dem Resource Interchange File Format (RIFF) basiert und zur Speicherung von digitalen Audiosignalen verwendet wird. Der Formattyp WAVE besteht dabei aus zwei Arten von Chunks – fmt und data chunk, wobei fmt chunk die Sample Rate und SampleBreite beschreibt und data chunk als Behälter für die Samples dient. Ein Sample stellt den Abtastwert dar, der während einer Analog-Digital-Wandlung ermittelt wird.

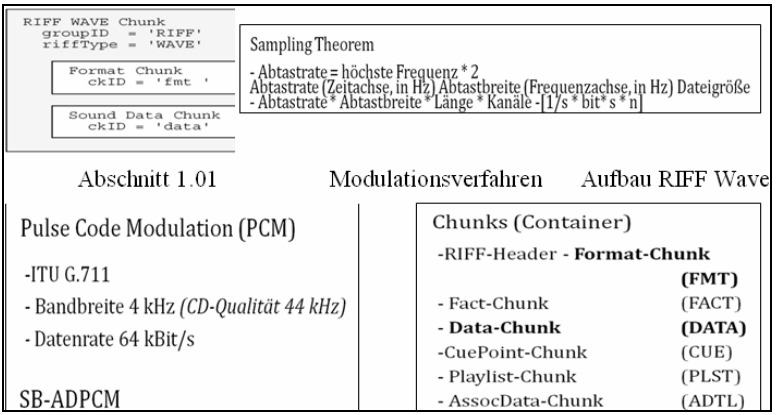


Abb. 59: Darstellung der Modulationsverfahren

Zur Illustration soll folgendes Rechenbeispiel dienen:

Pro Sekunde fallen Abtastrate x Bytes / Sample x Anzahl der Kanäle (mono = 1, stereo = 2) an Bytes an.

Beispiel: 1 Minute = 60 Sekunden, CD-Qualität (16 Bit = 2 Byte, 44.100 Hertz, stereo)

(44.100 Hz x 2 Byte x 2) x 60 s = 10.584.000 Byte = 10335,94 KiB = 10,09MiB

Zur Kompression der Audiodaten dient Advanced Audio Coding (AAC), das ein von der MPEG – Moving Picture Experts Group (Dolby, Fraunhofer-Institut für Integrierte Schaltungen in Erlangen, AT&T, Nokia und Sony), entwickeltes Audiodatenkompressionsverfahren darstellt, das als Weiterentwicklung von MPEG-2 Multichannel im MPEG-2-Standard spezifiziert wurde. Dieses Programm ist ein im Rahmen von MPEG-2 und MPEG-4 standardisiertes und von mehreren Firmen entwickeltes Verfahren, das mit faac auch über einen freien Encoder verfügt. AAC ist bei niedrigen Bitraten bis zu 160 kbit/s dem Format MP3 in der Klangqualität deutlich überlegen. Das Programm erlaubt die Nutzung eines Mehrkanal-Tons und wird von der Industrie wie zum Beispiel bei der Entwicklung von Handys und MP3-Playern breit zur Anwendung gebracht. Das Containerformat mp4 wurde neben dem Kompressionsschema definiert und erlaubt die Übermittlung von Metadaten (Tagging) oder die Verwendung eines Kopierschutzverfahrens (DRM). Für das Advanced Audio Coding sind insgesamt fünf Profile definiert, nämlich Low Complexity (LC), Low Delay (LD), Main Profile, High Efficiency (HE) und Scalable Sample Rate (SSR).

OGG/Vorbis Codec – Ogg Vorbis ist ein weiteres Audiokompressionsformat, das sich von MP3, VQF, AAC, und anderen digitalen Audioformaten dadurch unterscheidet, dass Ogg kostenlos ist und über keine Patentierung verfügt. Vorbis ist der Name eines spezifischen Audiokompres-

sionsschemas, dessen Aufbau und Struktur dem MPEG-4 Dateiformat MP4 ähneln, wobei als bekanntester Codec der Audio-Codec Vorbis fungiert. Es erhebt sich die Frage, wie gut Vorbis für eine Sprachaufnahme geeignet ist, wozu es festzuhalten gilt, dass es zwar über gewisse Qualitäten verfügt, keinesfalls jedoch eine optimale Lösung darstellt. Vorbis wurde hauptsächlich für die Audiokompression entwickelt, sodass spezielle Codecs wie z. B. Speex bei Sprachaufnahmen eine höhere Kompression als Vorbis erreichen.

L A M E A i n ' t a n M p 3 E n c o d e r (MPEG-1 Audio Layer 3, MP3) ist ein Dateiformat zur verlustbehafteten Audiodatenkompression. Das Format MP3 bedient sich dabei der Psychoakustik mit dem Ziel, nur für den Menschen bewusst hörbare Audiosignale zu speichern, womit dieses Format in qualitativer und funktioneller Hinsicht Formaten wie AAC (proprietär) oder Vorbis (kostenlos) unterlegen ist. Die Fraunhofer-Gesellschaft und andere Firmen besitzen Softwarepatente auf Teilverfahren, die für MPEG-Kodierung eingesetzt werden. Ein allumfassendes MP3-Patent gibt es jedoch nicht. Die Fraunhofer-Gesellschaft hat den größten Anteil an der Entwicklung des MP3-Standards und verfügt über Patente auf einige Verfahren zur MP3-Kodierung. Mit LAME können verschiedene Arten der Kodierung ausgewählt werden, so z. B. CBR (konstante Bitrate) und zwei Arten der variablen Bitrate, nämlich VBR (variable Bitrate) und ABR (durchschnittliche Bitrate).

Das Akronym F L A A C steht für Free Lossless Audio Codec und ist ein ähnliches Format wie mp3, das jedoch eine verlustfreie Audiodatenkompression besitzt. FLAAC ist der schnellste und am weitesten verbreitete verlustfreie Audio-Codec, der im Rahmen der Xiph.Org Foundation entwickelt wurde. Dieser Codec ist frei verfügbar und in seiner Nutzung nicht durch Softwarepatente beschränkt. Das Projekt besteht aus dem Streaming-Format, libFLAC (einer Bibliothek mit Referenz-Encoder und -Decoder und einer Metadaten-Schnittstelle), libFLAC++ (Objekt-Wrapper für libFLAC), flac (Kommandozeilentool zum Kodieren und Dekodieren von flac-Dateien mit libFLAC), metaflac (Kommandozeilentool zum Editieren der Metadaten von flac-Dateien) und Eingabefilter als Plugins für verschiedene Musik-Player (Winamp, X MultiMedia System usw.).

Die Programme l i b F L A C und l i b F L A C ++ sind unter einer angepassten Version der BSD-Lizenz verfügbar; flac, metaflac und die Plugins unter der General Product License. Im Gegensatz zu verlustbehafteten Audiodatenkompressionsverfahren wie MP3, AAC oder Vorbis ist die Komprimierung bei FLAAC verlustfrei. Die komprimierten Dateien sind um ein Vielfaches größer als bei verlustbehafteter Komprimierung wie z. B. in AAC oder MP3. FLAAC kann gestreamt werden und bietet Unterstützung für Mehrkanal-, Replay-, Gain- und Cuesheet. Außerdem können RIFF und AIFF Metadaten in FLAAC Dateien eingebunden werden. Angesichts dessen, dass es sich bei FLAAC um einen asymmetrischen Codec handelt, ist der Rechen-

aufwand für das Kodieren deutlich höher als jener für das Dekodieren. Das Format verfügt nur über eine Komplexitätsstufe, weshalb der Aufwand beim Dekodieren unabhängig von der Enkodereinstellung immer derselbe bleibt. Dies ist einer der Gründe, warum FLAAC auf Abspielgeräten eine beachtliche Verbreitung erlangt hat. Sein durchschnittlicher Kompressionsfaktor liegt bei rund 0,5, wobei die Größe der Ausgangsdatei durch die Kodierung auf die Hälfte reduziert wird. Die Kompression besteht aus acht Abstufungen (8 = maximal, 5 = Standard), und die Unterschiede hinsichtlich Größe sind minimal. Die zur Komprimierung benötigte Zeit wächst überproportional mit der Kompressionsstufe, wobei das Original im Zuge des Verfahrens auf durchschnittlich etwa 60% der Ausgangsgröße reduziert wird.

Robert Thomann (Graz)

Das Gralis-Anketarium

18. Mithilfe des online aufrufbaren Programms Anketarium konnte im Rahmen des Projektes „Die Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen“ die Möglichkeit geschaffen werden, Fragebögen, Umfragen und Datenerhebungen online durchzuführen, wobei es allen Personen mit individuell granulierbaren Zugangsberechtigungen offen steht, Fragebögen und andere Dokumente online zu erstellen und die auf einem Webserver gespeicherten Ergebnisse jederzeit abzurufen.

Das Programm Anketarium wurde im Zeitraum zwischen Februar und August 2007 entwickelt und zeichnet sich dadurch aus, das es allen an der Erstellung von Online-Fragebögen und Umfragen interessierten Personen ermöglicht, diese problemlos, effizient und den eigenen Bedürfnissen entsprechend anzulegen. Die Bedienung des Anketariums erfolgt über einen Browser wie etwa den Internet Explorer oder Mozilla Firefox; zusätzliche Software wird nicht benötigt. Um der Anforderung nach Mehrsprachigkeit gerecht werden zu können wurden die Bedienelemente viersprachig (BKS, deutsch, russisch und slowenisch) ausgeführt. Des Weiteren hat jede das Anketarium nutzende Person die Möglichkeit, eigene Umfragen in eine andere der genannten Sprachen zu übersetzen und auf diese Weise mehrsprachige Umfragen zu erstellen. Liegt der Fragebogen mehrsprachig vor, so können die ProbandInnen ihre bevorzugte Sprache aus einer Auswahlliste wählen und den Fragebogen in dieser Sprache angezeigt bekommen.

Die inhaltliche Struktur des Anketariums stellt sich dreigliedrig dar, wobei im Hauptmenü eine Unterscheidung zwischen wissenschaftlichen Umfragen (Wissenschaftliche Umfragen), Umfragen für Zwecke des Unterrichts (Edukative Umfragen) und Umfragen unterschiedlichen Inhalts (z. B. Fragen des Monats –Andere Umfragen) vorgenommen wird.

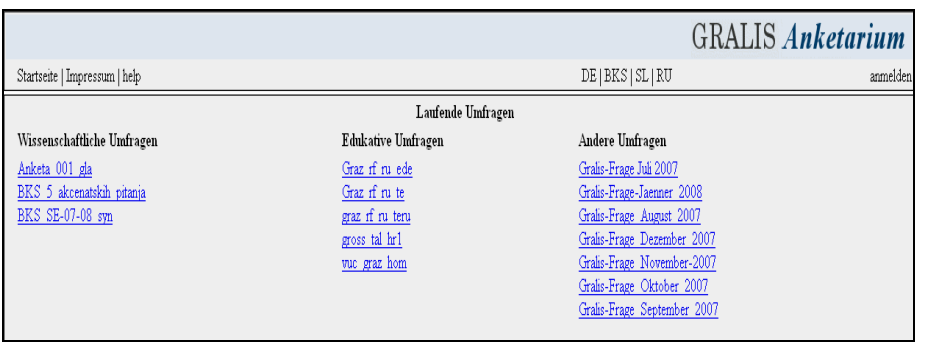


Abb. 60: Die Startseite des Gralis-Anketarium

Umfragen erstellen: Anmeldung. Um Umfragen erstellen zu können, muss man sich zuerst anmelden, wozu ein Betätigen des Anmeldeknopfes rechts oben im Browserfenster erforderlich ist:

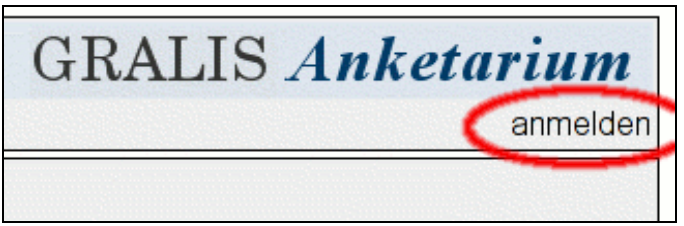


Abb. 61: Die Anmeldung im Gralis-Anketarium

In einem nächsten Schritt werden sodann die Benutzerdaten eingegeben, und es erfolgt ein Klick auf den Befehl „anmelden“.

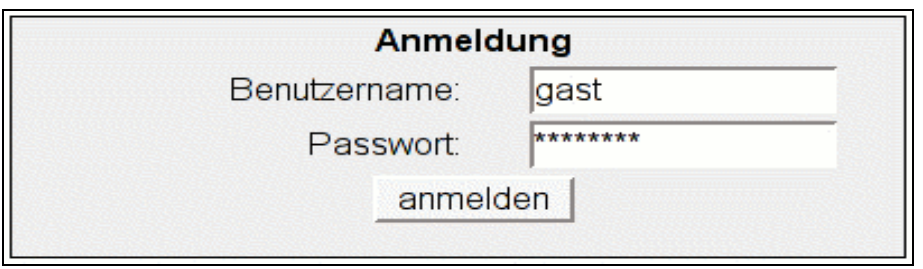


Abb. 62: Durchführen der Anmeldung

Nach Durchführen der Anmeldung, steht links oben unten dargestelltes das Navigationsmenü zur Verfügung, das im Folgenden eingehender erläutert wird.

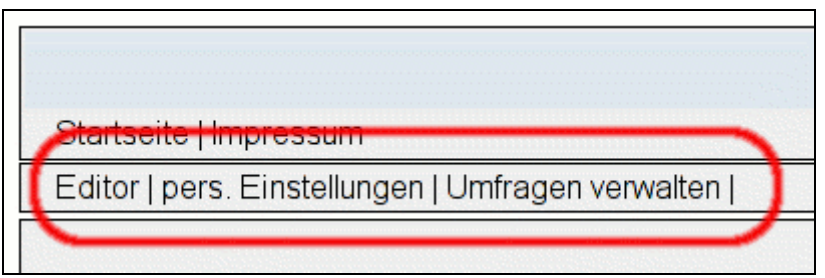


Abb. 63: Navigationsmenü

Nach der Anmeldung befindet man sich in der Umfragenverwaltung, die neben verschiedenen Verwaltungsaufgaben auch die Möglichkeit bietet, an dieser Stelle eine neue Umfrage zu erstellen. Dazu klickt man auf den Knopf mit der Aufschrift „Neue Umfrage anlegen“, woraufhin man an die entsprechende Editor-Funktion weitergeleitet wird.

Der Editor. Ruft man den Editor über den Navigationsmenüeintrag „Editor“ auf, wird zunächst keine Umfrage geladen oder erstellt. Um

diesen Schritt durchzuführen, wählt man aus der Liste rechts im Bild den Eintrag „neue Umfrage“ und klickt anschließend auf „OK“.

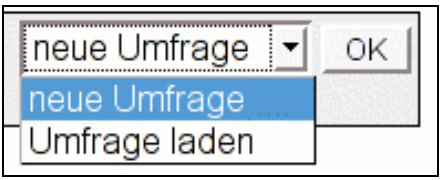


Abb. 64: Neue Umfragen erstellen mithilfe des Editors

In weiterer Folge füllt man das anschießend erscheinende Formular entsprechend aus und klickt daraufhin auf den Befehl „anlegen“. Der Name der Umfrage dient der eindeutigen Identifikation der Umfrage und darf keinerlei Sonderzeichen enthalten. Eine andere Möglichkeit, zu diesem Formular zu gelangen, liegt darin, die neue Umfrage durch Betätigen des diesbezüglichen Befehls in der Umfragenverwaltung zu erstellen.

Abb. 65: Erstellen einer neuen Umfrage

Nach Abschluss dieses Arbeitsschrittes wurde die Struktur für eine neue Umfrage erstellt. In weiterer Folge bietet das System die Möglichkeit, Umfragen in Gruppen von Fragen zu gliedern, wozu der Befehl „neue Gruppe anlegen“ anzuwählen und die Eingabe mit „OK“ zu bestätigen ist.

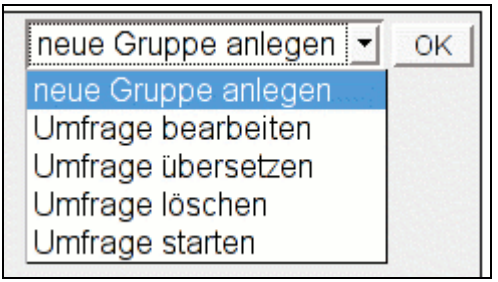


Abb. 66: Anlegen einer neuen Gruppe

Nachdem man das sodann erscheinende Formular ausgefüllt und bestätigt hat, wurde eine Gruppe angelegt, die nun mit Fragen befüllt werden kann.

Abb. 67: Benennung einer neuen Gruppe

Bei der Erstellung von Fragen geht man analog zur Anlegung einer Gruppe vor, wobei man eine „neue Frage“ aus der hinzugekommenen Auswahlliste auswählt und sodann auf „OK“ klickt. Nach Ausfüllen des daraufhin erscheinenden Formulars kann der entsprechende Antwortmodus gewählt werden.

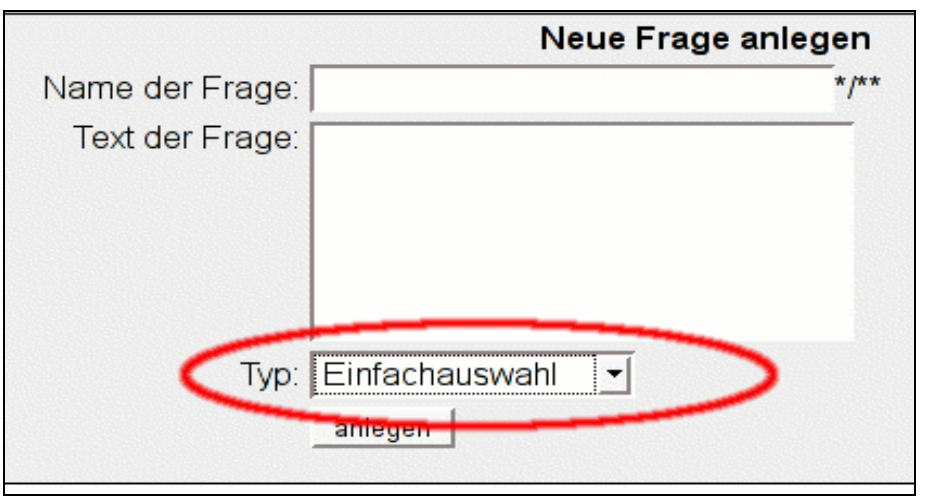


Abb. 68: Anlegen einer neuen Frage

In Abhängigkeit von der Entscheidung für eine „Einfachauswahl“ oder eine „Mehrfachauswahl“ besteht die Möglichkeit, individuell bevorzugte

Antwortmöglichkeiten zu erstellen. Hierzu wählt man den Befehl „neue Antwort“ aus der Entsprechenden Auswahlliste aus und klickt auf „OK“.

In unten dargestellter Abb. 69 sieht man eine Beispielfrage, die mit zwei Antworten versehen wurde, wobei der Inhalt von „Wert“ in der Datenbank abgelegt wird und für die spätere Auswertung von Bedeutung ist. Unter „Text“ findet man die am Bildschirm angezeigte Antwortmöglichkeit für sämtliche befragte Personen.

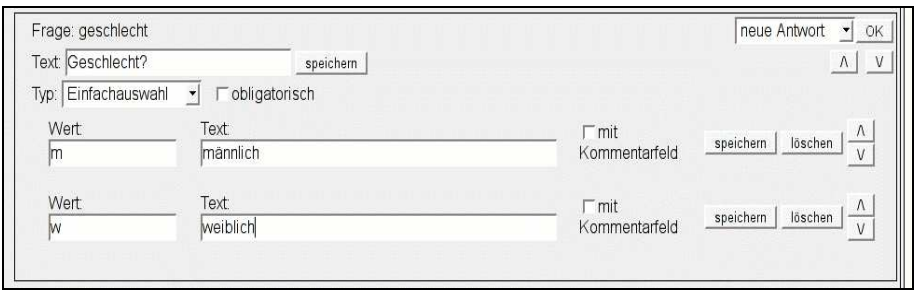


Abb. 69: Muster einer Frage mit zwei Antwortmöglichkeiten

Beim Ausfüllen des Fragebogens stellt sich diese wie folgt am Bildschirm dar:

Abb. 70: Arbeitsansicht für eine ausfüllende Person

Umfragenverwaltung: Mehrsprachige Umfragen. Es wird generell empfohlen, eine Umfrage zuerst in einer Sprache fertig zu stellen und die abgeschlossene und für eine Aussendung vorgesehene Umfrage erst in einem weiteren Arbeitsschritt in eine andere Sprache zu übersetzen. Dazu ist zuallererst ein Klick auf den im Navigationsmenü dargestellten Befehl „Umfragen verwalten“ vorzunehmen, um auf diese Weise zur Umfragenverwaltung zu gelangen. In weiterer Folge ist der in Abb. 71 blau unterlegte Menüpunkt „übersetzen“ anzuwählen.

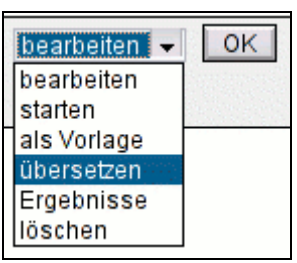


Abb. 71: Übersetzen einer Umfrage

Nun können die zu übersetzenden Textpassagen in die auf der rechten Seite dargestellten Textfelder eingetragen werden:

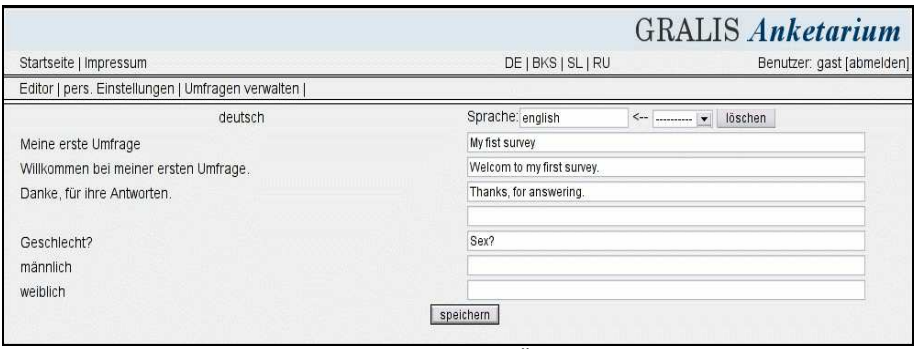


Abb. 72: Eintragen der Übersetzung

Es sei angemerkt, dass das Übersetzungswerkzeug neben seiner eigentlichen Funktion auch dazu dient, Fehler oder abzuändernde Textstellen in jeder beliebigen Umfrage auszubessern, wozu in der Auswahlliste die gewünschte Sprache anzuwählen ist, woraufhin Fehler ausgebessert bzw. Änderungen vorgenommen werden können.

E r g e b n i s s e u n d S t a t i s t i k . Wird eine Umfrage bzw. ein Fragebogen als von einer ausreichenden Zahl als beantwortet und abgeschlossen erachtet, liegt der nächste Schritt darin, die Ergebnisanzeige zu wählen, wozu der Befehl „Ergebnisse“ zu aktivieren und mit „OK“ zu bestätigen ist. Abb. 73 zeigt eine beispielhafte Anzeige der Ergebnisse einer abgeschlossenen Umfrage.

Abb. 73: Muster der Darstellung von Ergebnissen einer Umfrage

Eine exakte Darstellung der in obiger Abbildung einzusehenden Ergebnisse bietet sich unter dem Menüpunkt „Statistik“, nach dessen Anwahl man eine detaillierte Auflistung in absoluten Zahlen und perzentuellen Anteilen erhält.

Antwort	Anzahl	%
Begeisterung	23	79.31%
Familie	7	24.14%
Berufsausbildung	8	27.59%
Literatur	6	20.69%
Translation	2	6.9%
Filme und Musik	0	0%
andere	4	13.79%

Abb. 74: Muster der Anzeige im Menüpunkt „Statistik“

Das Programm Anketarium erwies sich bereits ab dem Zeitpunkt seiner Fertigstellung und nach dem Verfassen der ersten Testeingaben als überaus hilfreiches Werkzeug für die Erstellung und Auswertung von Umfragen oder Fragebögen, dessen Anwendungsgebiet neben wissenschaftlichen Datenerhebungen und Auswertungen auch vor allem im edukativen Sektor als Unterstützung beim Verfassen studentischer Arbeiten beheimatet sein wird.

Stefan Kofler – Arno Wonisch (Graz)

Das Gralis-Rezensarium

19. Das von Stefan Kofler im Rahmen des Projektes „Die Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen“ entwickelte Programm mit dem Titel Rezensarium dient dazu, Rezensionen bzw. Änderungsvorschläge zu wissenschaftlichen Aufsätzen bzw. in weiterer Folge auch zu Dokumenten jeglicher Art online zu erstellen und den jeweiligen AutorInnen zukommen zu lassen. Zu diesem Zweck ist es erforderlich, dass sich im Vorhinein ausgewählte RezensentInnen mit individuellen Passwörtern ins Programm einloggen, die Aufsätze der ihnen zugeteilten AutorInnen einsehen, eventuelle Verbesserungsvorschläge einbringen und die Dokumente sodann für die Publikation freigeben. Die gesamte Benutzeroberfläche des Rezensariums ist zweisprachig (deutsch und BKS).

Für seitens der Projektverantwortlichen für die Erstellung von Rezensionen ausgewählte Personen haben als ersten Arbeitsschritt das Einloggen vorzunehmen, wobei als Benutzername die E-Mail-Adresse dient und das Passwort von den Projektverantwortlichen bzw. Administratoren an die einzelnen RezensentInnen verschickt wird. Dieser Arbeitsschritt des Logins wird in Abb. 75 dargestellt.

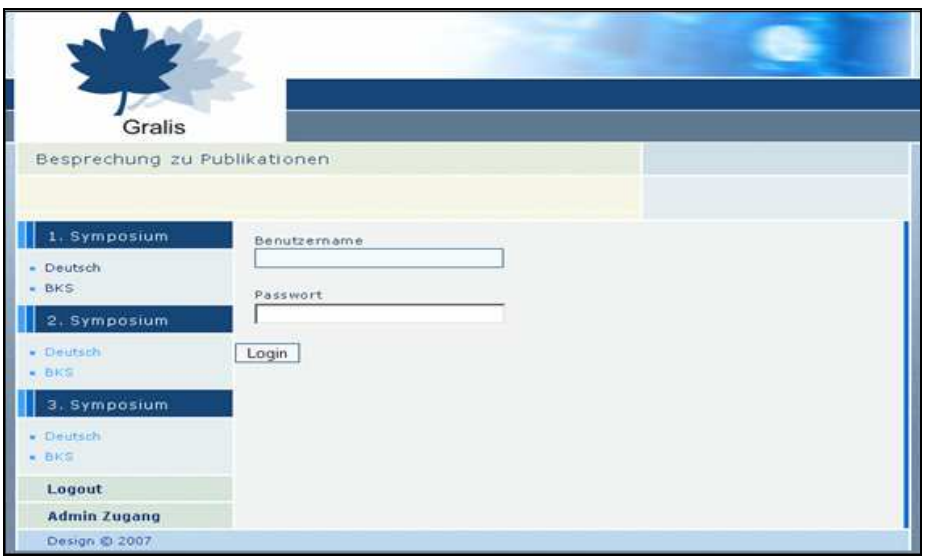


Abb. 75: Das Einloggen

In einem zweiten Arbeitsschritt, der in unten stehender Abb. 76 gezeigt wird, bietet sich einer erstmals ins Programm einsteigenden Person die Möglichkeit, die persönlichen Daten (inklusive Bankverbindung zur Auszahlung eines eventuell vorgesehenen Honorars) einzugeben oder/und sodann die Arbeiten zugeteilter AutorInnen zu begutachten.

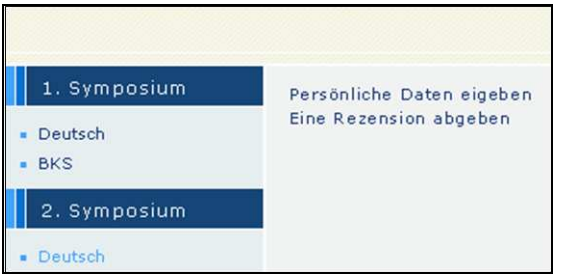


Abb. 76: Auswahl zwischen persönlichen Daten eingeben und Verfassen einer Rezension

Es wird auf alle Fälle empfohlen, persönliche Daten unverzüglich beim erstmaligen Einloggen einzutragen, weil diese Daten 1) ausschließlich den Projektverantwortlichen zugänglich sind und sie 2) ab der ersten Rezension für alle darauf folgenden gespeichert bleiben. Die Bestätigung der Eingabe erfolgt durch einen Klick auf „OK“.

Hat man diesen Arbeitsschritt durchgeführt und entscheidet sich in weiterer Folge dafür, mit dem Verfassen einer Rezension zu beginnen, so öffnet sich nach Betätigen des in Abb. 76 dargestellten Befehles „Eine Rezension abgeben“ eine Maske, aus der hervorgeht, welche Aufsätze bzw. Beiträge von den Projektverantwortlichen für eine Rezensierung ausgewählt wurden.

Aus Abb. 77 geht hervor, dass hier gewähltem Rezensenten insgesamt ein Beitrag für eine Rezensierung zugeteilt wurde, die in diesem Falle bereits erfolgt ist. Dies lässt sich an dem Eintrag in blauer Farbe in der Rubrik „Bereits bewertet“ erkennen. Durch einen Klick auf die Bezeichnung des Dokumentes in rechter Spalte lässt sich dieses öffnen, wogegen eine Anwahl des Namens des angeführten Autors in linker Spalte die Eingabemaske für den Eintrag von Änderungsvorschlägen und Bemerkungen aktiviert.

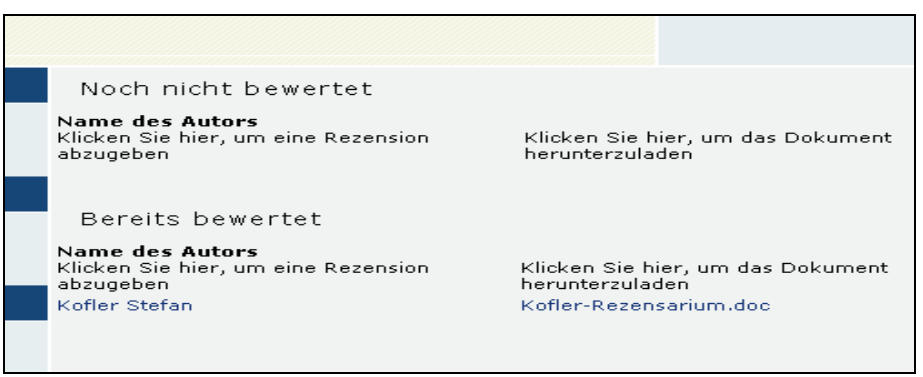


Abb. 77: Übersicht über die zugeteilten Rezensionen und den Stand der Bewertungen

Jeder Verbesserungsvorschlag verfügt über eine Überschrift, die z. B. Seite, Kapitel, Zeile, Nr. u. Ä. lauten kann, weiters über den Originaltext in linker Spalte und über die von den rezensierenden Personen einzutragenden

Korrekturvorschläge, Bemerkungen oder Anregungen in der Spalte auf der rechten Seite der Maske. Siehe dazu die unten stehende Darstellung dieser Maske, die über insgesamt zehn Felder – sowohl für den Originaltext als auch für den verbesserten Wortlaut umfasst.

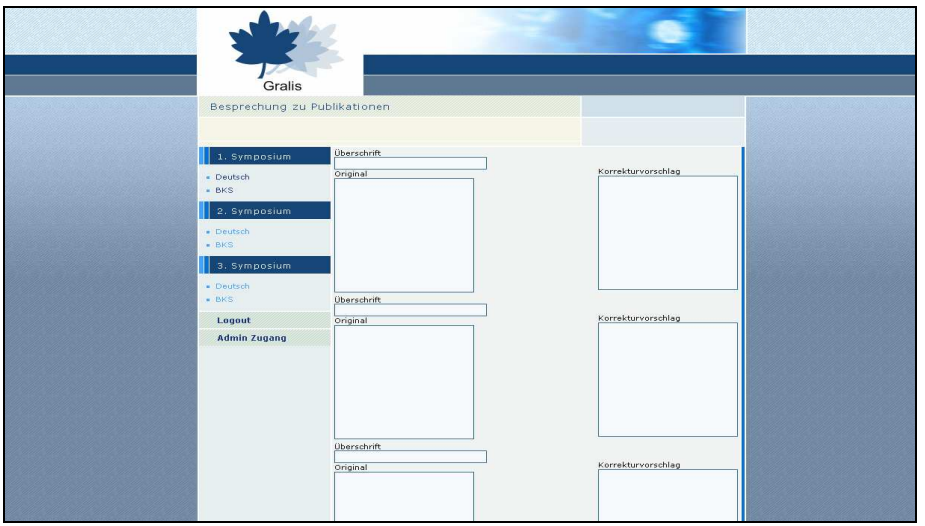


Abb. 78: Eingabemaske für die Rezension

Sollte man zu dem für eine Rezensierung zugeteilten Text keinerlei Bemerkungen tätigen möchten, kann man dies durch einen diesbezüglichen Kommentar im ersten für einen Korrekturvorschlag vorgesehenen Fenster tun.

Nach dem Eintragen aller Vorschläge und Bemerkungen seitens der rezensierenden Person wird in einem finalen Arbeitsschritt darum ersucht, ob der eben rezensierte Beitrag bzw. das Dokument für eine Veröffentlichung empfohlen werden kann. Dazu stehen die in Abb. 79. abgedruckten drei Optionen zur Verfügung, die 1) eine vorbehaltlose Veröffentlichung dieses Beitrages empfehlen, 2) eine Publikation nach Vornahme von Änderungen nahelegen oder 3) von einer Veröffentlichung des eingesehenen Beitrages abraten.

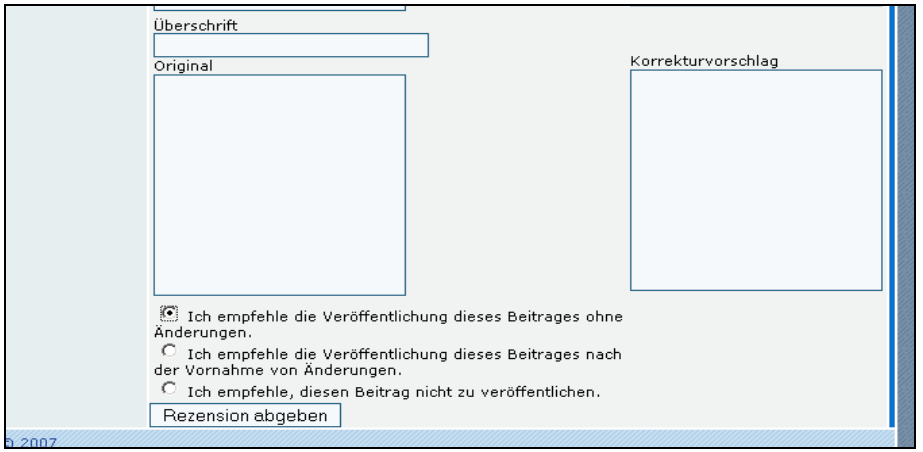


Abb. 79: Auszuwählende Meinung betreffend die Veröffentlichung eines Beitrages

Durch einen Klick auf den ebenfalls in oben stehender Abbildung dargestellten Befehl „Rezension abgeben“ werden sämtliche Einträge der rezensierenden Person per E-Mail zeitgleich an die Autorin/den Autor und an die Projektverantwortlichen übermittelt. Es ist nun Aufgabe der Autorin/des Autors, die zugesandten Änderungsvorschläge in den Text einzuarbeiten und diesen sodann den Projektverantwortlichen zukommen zu lassen.

In technischer Hinsicht basiert das Rezensarium auf dem Open-Source-Programm Ruby on Rails und einer MySQL-Datenbank, wobei von den Projektverantwortlichen, die zugleich auch als Administratoren fungieren, jederzeit neue RezensentInnen und AutorInnen hinzugefügt werden können. Seinen ersten „Einsatz“ erfuhr das Rezensarium bei der Bewertung der in diesem Band abgedruckten Beiträge des ersten Symposiums im Rahmen des Projektes „Die Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen“, wobei insgesamt 79 Gutachten zu Aufsätzen verfasst wurden.

Branko Tošović (Graz)

Das Gralis-Bibliothekarium

20. Das Gralis-Bibliothekarium stellt einen Teil des Gralis-Komplementariums dar und dient als bibliographische Ergänzung zum Gralis-Korpus sowie als Hilfsmittel bei der Durchführung wissenschaftlicher Projekte und im Unterricht. Seine Aufgabe liegt nicht nur in der Sammlung, Bearbeitung und schnellen Auffindung bibliographischer Angaben, sondern auch in der Verwaltung von Informationen im Interesse einer möglichst problemlosen und schnellen Zitierung, wodurch die Erstellung von Literaturverzeichnissen in Büchern, Sammelbänden, Artikeln u. a. erheblich erleichtert wird. Das Bibliothekarium besteht aus drei Teilen, von denen der erste Publikationen in lateinischer Schrift umfasst (Lat-Bibliothekarium), der zweite für kyrillische Werke vorgesehen ist (Cyr-Bibliothekarium) und der dritte bibliographische Angaben des Leiters des Gralis-Korpus enthält (BT-Biblio). Im Sinne einer Vereinfachung der Suchabfragen werden die kyrillischen bibliographischen Einheiten für das Serbische ins Lat-Bibliothekarium integriert, sodass das Cyr-Bibliothekarium aus Literatur für die Sprachen bulgarisch, mazedonisch, russisch, ukrainisch und weißrussisch besteht. Zum Zeitpunkt der Drucklegung dieses Bandes enthielt das Lat-Bibliothekarium 1.372 Einträge und das Cyr-Bibliothekarium 2.449 bibliographische Einheiten.

Abb. 80: Das Startinterface des Gralis-Bibliothekariums

Der Zutritt zum Administratoren-Bereich des Bilbiothekariums ist im Gegensatz zur Suche nur für registrierte BenutzerInnen möglich, wobei solche jederzeit vom Administrator flexibel angelegt werden können.

Abb. 81: Das Einloggen ins Gralis-Bibliothekarium

Die bibliographische Struktur zur Eingabe von Angaben ist in die Rubriken „Titeldaten“, „Verlagsangaben“ und „weitere Felder“ unterteilt. In die Rubrik „Titeldaten“ werden folgende Angaben eingetragen: Zitierung (Art der Zitierung, die zu Beginn der bibliographischen Einheit aufscheint), Herausgeber, Titel und (falls vorhanden) Untertitel. Handelt es sich um eine Übersetzung, sind auch die Menüpunkte „Originaltitel“, „Übersetzung“, „Sprache des Originals“, „Typ“ (z. B. Buch, Lehrbuch, Handbuch, Sammelband, Zeitschrift u. a.) und Genre auszufüllen. Sämtliche Einträge können auf einfache Weise dupliziert werden, was etwa im Falle von Reihen-, Serien- oder Zeitschriftentitel eine erhebliche Vereinfachung darstellt.

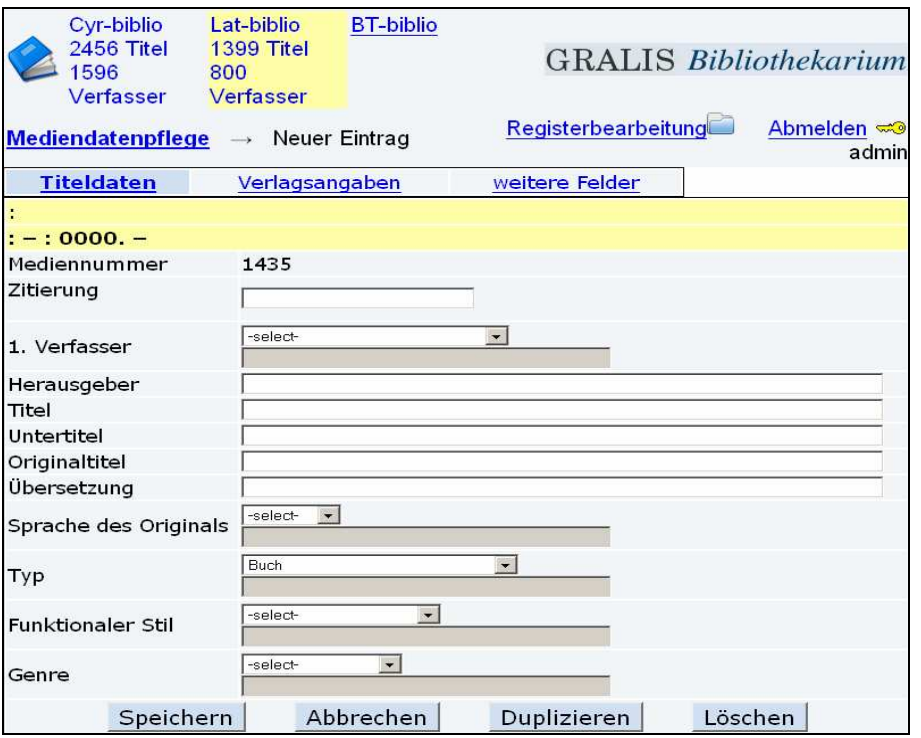


Abb. 82: Eingabefenster des Gralis-Bibliothekariums

Im Zuge des Ausfüllens werden in der Datenbank bereits vorhandene Einträge automatisch angezeigt, wodurch ein erneutes Eingeben gleicher Angaben (z. B. Namen von AutorInnen) entfallen kann.

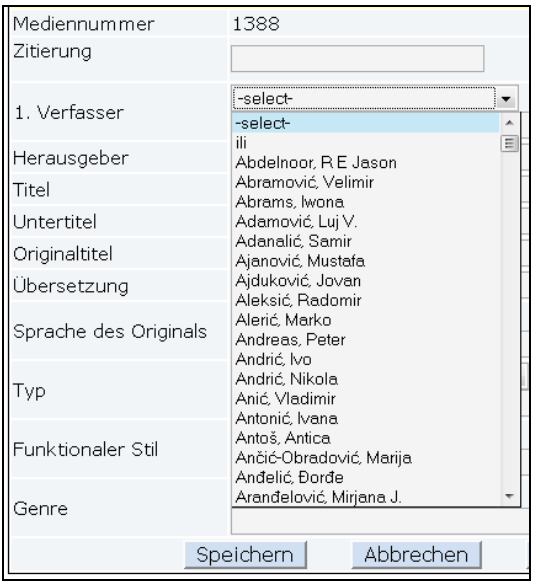


Abb. 83: Eingabe von Verfassernamen mit bereits vorhandenen Einträgen

Die Rubrik „Verlagsangaben“ umfasst folgende Informationen: Verlagsort, Verlag, Erscheinungsjahr, Umfang, Sprache, Quelle (Zeitschrift, Sammelband), Reihe (Serie), Auflage, Zitierungsstandard (angeboten werden Standards für das Deutsche, für BKS, russisch, englisch wie auch ein modifizierter Gralis-Standard), Schrift und zusätzliches Medium.

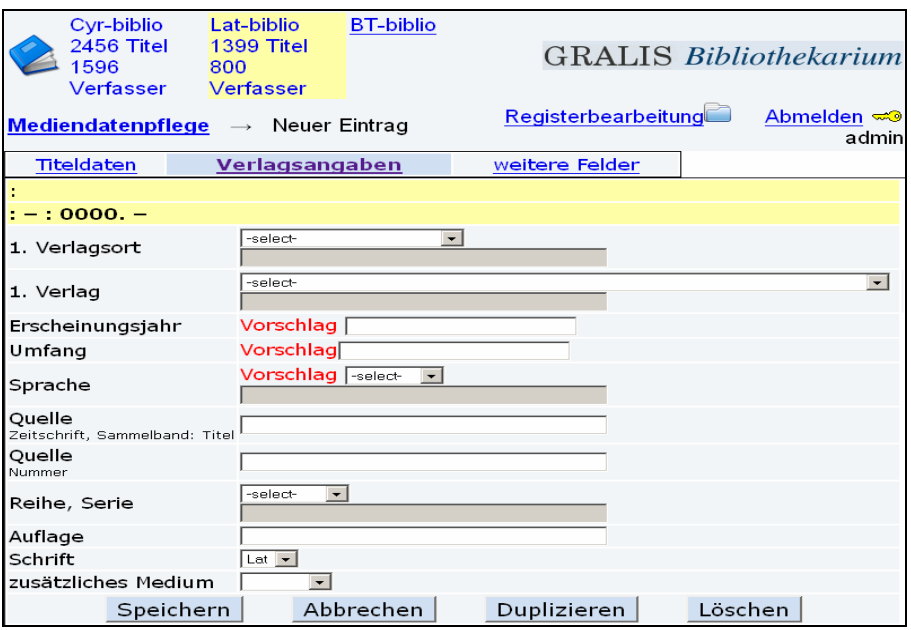


Abb. 84: Eingabe von Verlagsangaben

In der Rubrik „weitere Felder“ ergehen Informationen zu „Biblio Quelle“ (z. B. einer Bibliographie in einer Zeitschrift, einer Privat- oder Universitätsbibliothek u. a.), Schlagwörtern, Bemerkungen, Operator(en) [d. h. bearbeitenden Personen], Status (unbearbeitet, im Bearbeitung, fertig bearbeitet), Katalogisierungsdatum und zur letzten Änderung.

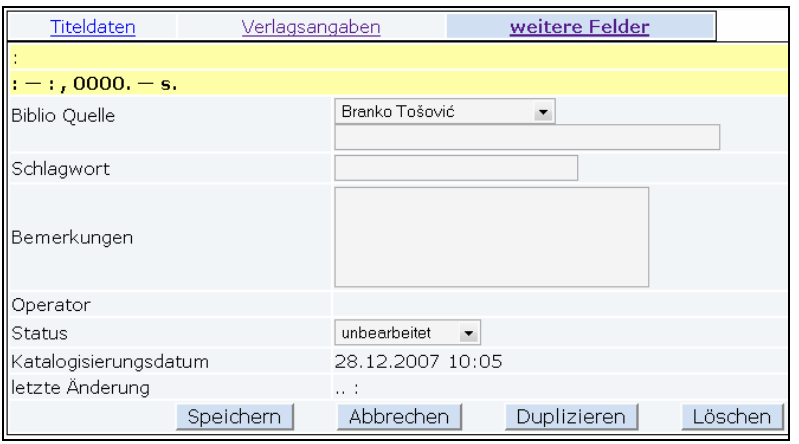


Abb. 85: Eingabe von „weiteren Feldern“

Eine Suchabfrage im Gralis-Bibliothekarium beginnt mit der Wahl des Lat- oder Cyr-Bibliothekariums, wobei mehrere Suchoptionen zur Auswahl stehen (Suchkriterium 1, Suchkriterium 2...).

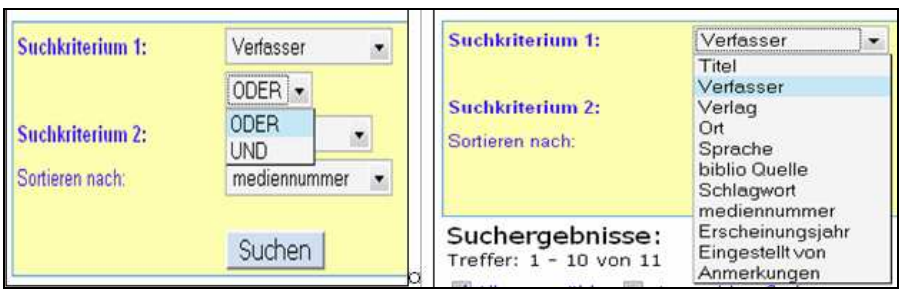


Abb. 86: Wahl von Suchkriterien

Betrifft eine Suchabfrage ein einziges Wort, so muss hinter diesem ein Prozentsymbol (%) gesetzt werden. Generell sind Abfragen zu allen Rubriken möglich, wobei als Ergebnis Informationen zur Evidenznummer, AutorIn, Titel und Jahr der Herausgabe aufscheinen.

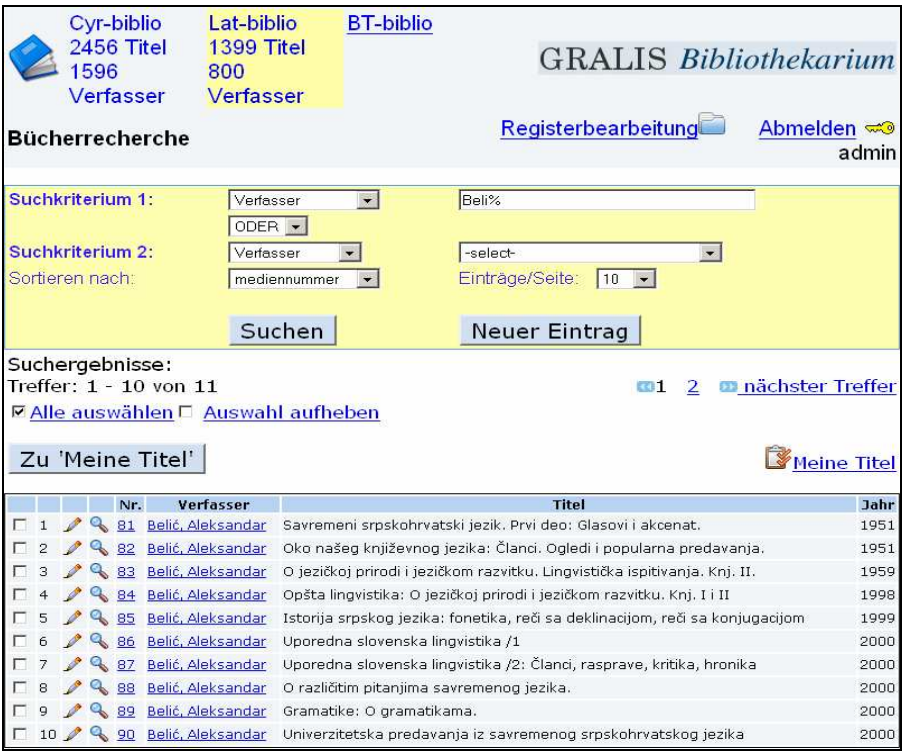


Abb. 87: Ergebnis einer Suchabfrage

Die gesamte Information zu einer gesuchten bibliographischen Einheit erhält man durch einen Klick auf die auf der linken Seite des AutorInnennamens verzeichnete Evidenznummer, auf das Lupensymbol oder auf den Namen der/des Autor(In).

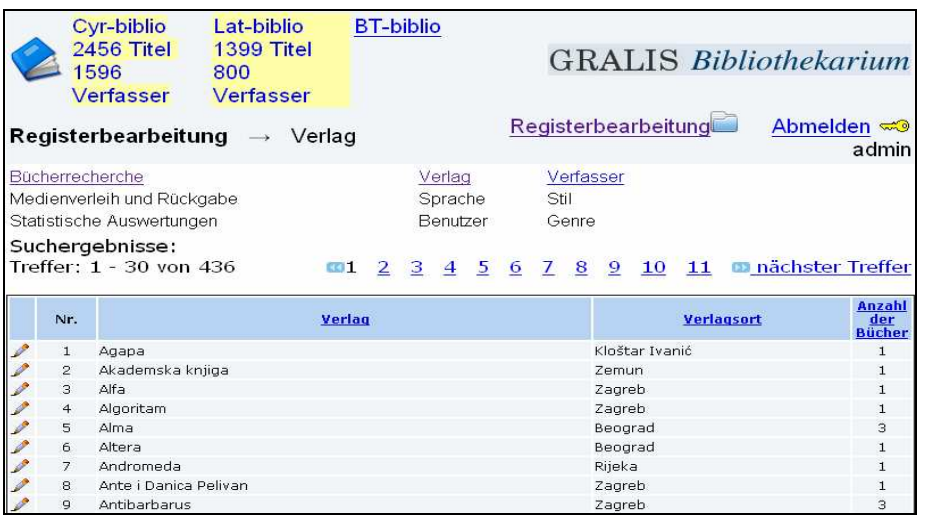
Eine weitere, den Arbeitsprozess beschleunigende Funktion des Gralis-Bibliothekariums liegt darin, dass eine individuelle Auswahl von Buchtiteln vorgenommen werden kann. Dies geschieht durch ein Anwählen der gewünschten Titel in der Spalte am linken Rand der oben dargestellten Abbildung und durch einen Klick auf „zu ʻMeine Titel'“ (linke Bildmitte). Nach einer Anwahl von „Meine Titel“ (gegenüberliegend am rechten Bildrand) folgt eine Vorschau über die gewählten Titel, die sodann mittels Befehl „drucken“ in eine zitierfertige Form gebracht werden können.

Belić 1951: Belić, Aleksandar. Savremeni srpskohrvatski jezik. Prvi deo: Glasovi i akcenat. — Beograd: Naučna knjiga, 1951. — 181 s.	
Verfasser:	**Belić, Aleksandar**
Herausgeber	
Titel	Savremeni srpskohrvatski jezik. Prvi deo: Glasovi i akcenat.
Untertitel	
Originaltitel	
Übersetzung	
Sprache des Originals	
Verlag	Naučna knjiga
Verlagsort	Beograd
Jahr	1951
Umfang	181
Sprache	SR
Chiffre	S-
Quelle	
Reihe	
Auflage	
Standard	
Schrift	Lat
zusätzliches Medium	
Funktionaler Stil	
Genre	
Typ	Buch
Zitierung	Belić 1951:
Biblio Quelle	Branko Tošović
Schlagwort	
Bemerkungen	
Katalogisierungsdatum	01.11.2007 12:00
letzte Änderung	30.11.2007 21:23

Abb. 88: Darstellung der gesamten erfassten Information zu einer Publikation

Im oberen, gelb markierten Teil von Abb. 88 ist eine Darstellung einer Zitatangabe gemäß dem gewählten Standard zu sehen.

Sollte die Darstellung einer Statistik nach selbst zu wählenden Kriterien gewünscht werden, kann dies durch ein Öffnen der Rubrik „Registerbearbeitung“ getan werden. Siehe dazu Abb. 88.



	Nr.	Verlag	Verlagsort	Anzahl der Bücher
	1	Agapa	Kloštar Ivanić	1
	2	Akademska knjiga	Zemun	1
	3	Alfa	Zagreb	1
	4	Algoritam	Zagreb	1
	5	Alma	Beograd	3
	6	Altera	Beograd	1
	7	Andromeda	Rijeka	1
	8	Ante i Danica Pelivan	Zagreb	1
	9	Antibarbarus	Zagreb	3

Abb. 89: Die Statistik des Gralis-Bibliothekariums

Arno Wonisch (Graz)

Das Gralis-Personalium

21. Eine schnelle, effiziente und vor allem einfach zu handhabende Verwaltung von größeren Mengen an Datenmaterial stellt eine der Grundvoraussetzungen für eine rasche und schlanke Projektadministration dar. Diesen Ansprüchen Genüge getan werden kann durch eine vereinheitlichte und zentrale Sammlung wesentlicher Datenquellen, die im Sinne einer kompakten und klar gegliederten Verwaltung für die administrierenden Personen möglichst auf Knopfdruck abrufbar sein sollten.

Ausgehend vom Wunsch nach konkreter Umsetzung dieser Erfordernisse stellte sich auch zu Beginn des Projektes „Die Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen“ die Frage, wie der zu erwartende administrative Aufwand nach Möglichkeit gering gehalten und in hohem Maße automatisiert werden könnte. Aus diesem Grund begannen bereits geraume Zeit vor dem offiziellen Projektstart im Oktober 2006 diesbezügliche Überlegungen, denn die sukzessive einsetzende und stetig zunehmende Kommunikation mit über 100 kontaktierten und am Projekt interessierten WissenschaftlerInnen aus über zehn Staaten sollte in überschaubare Bahnen gelenkt werden.

Der erste konkrete Arbeitsschritt betraf deshalb die Erfassung elementarer Daten all jener Personen, die sich nach ausführlichen Informationen über die Konzeption und Abwicklung des Projektes seitens der Projektleitung definitiv dazu bereit erklärten, an unserem dreijährigen Forschungsvorhaben mitzuwirken und dies auch schriftlich bestätigten. Dank der Mithilfe von Herrn Dieter Schicker (Institut für Informationsverarbeitung in den Geisteswissenschaften der Karl-Franzens-Universität Graz – INIG) kam es sodann bereits im Herbst 2005 – ein Jahr vor dem geplanten Projektstart – zur Ausarbeitung der so genannten „A n m e l d u n g – P r i j a v a / P r i j a v n i c a“, die sich auf der Gralis-Website in der Rubrik Projektarium befand. Nach Aktivieren des gleichnamigen Links öffnete sich ein Fragebogen, der neben persönlichen Angaben auch Fragen zur wissenschaftlichen und edukativen Tätigkeit sowie zu den Modalitäten der gewünschten Projektmitarbeit beinhaltete. Auf diese Weise konnte ein Medium geschaffen werden, dass die für die Projektleitung wesentlichen Angaben an einem Ort versammelte.

Diese personelle Evidenz erwies sich in der Anfangsphase des Projektes auch durchaus als zweckmäßig, denn diverse personenbezogene Informationen (z. B. Wohnadresse oder Publikationsliste) konnten in kürzester Zeit abgerufen werden. Mit zunehmendem Projektfortgang und weiterem Anwachsen der Informationsmengen sollten jedoch schon bald kleinere Nachteile dieser Datenbank zu Tage treten, die einerseits darin lagen, dass (1) vorgenommene Einträge nicht (bzw. nur mit größerem Aufwand) abgeändert werden konnten, (2) der Fragebogen auch Fragen beinhaltete, die sich als nicht

primär wichtig herausstellten und (3) kein Hinzufügen weiterer Applikationen (Eingabe von Text- oder Audiodateien) möglich war.

Aus diesen Gründen wurde schließlich im Verlaufe des Sommers 2007 die Idee geboren, die Erfassung personenbezogener Daten einerseits zu straffen und sie andererseits auszuweiten und multifunktional zu gestalten. In Zusammenarbeit mit Frau Olga Lehner, die als Projektmitarbeiterin und Programmiererin auch für einige weitere wesentliche Programme im Rahmen des Projektes verantwortlich zeichnet, begann schließlich die Entwicklung des so genannten Gralis-Personaliums (basierend auf einer MySql-Datenbankstruktur), das eine Weiterentwicklung der hier kurz beschriebenen Anmeldungs-Datenbank darstellt und über eine BenutzerInnen- sowie eine Administrationsoberfläche verfügt.

Das Gralis-Personalium für BenutzerInnen. Mit der Fertigstellung des Gralis-Personaliums besteht nun seitens aller BenutzerInnen die Möglichkeit, sich in einer zentral verwalteten Online-Datenbank für eine Projektmitarbeit anzumelden und die eigenen Einträge und Angaben persönlich zu verwalten sowie im Bedarfsfalle abzuändern. Das Gralis-Personalium besitzt eine zweisprachige Menüführung – deutsch und BKS –, wobei eine Erweiterung um zusätzliche Sprachen durch Hinzufügen einer Sprachwahlleiste jederzeit möglich ist.

Nach Anwahl des entsprechenden Links auf der Gralis-Website (Rubrik „Projektarium“) treffen Benutzende zu Beginn auf unten stehendes Fenster, wobei bei einem erstmaligen Besuch die Aufmerksamkeit der in der unteren Bildhälfte dargestellten „Anmeldung – Prijava/Prijavnica“ samt darauf folgendem Fragebogen gelten soll. Das in der Bildmitte abgebildete Loginkästchen dient für bereits registrierte BenutzerInnen und ist erst bei einem neuerlichen Aufsuchen der Seite auszufüllen.

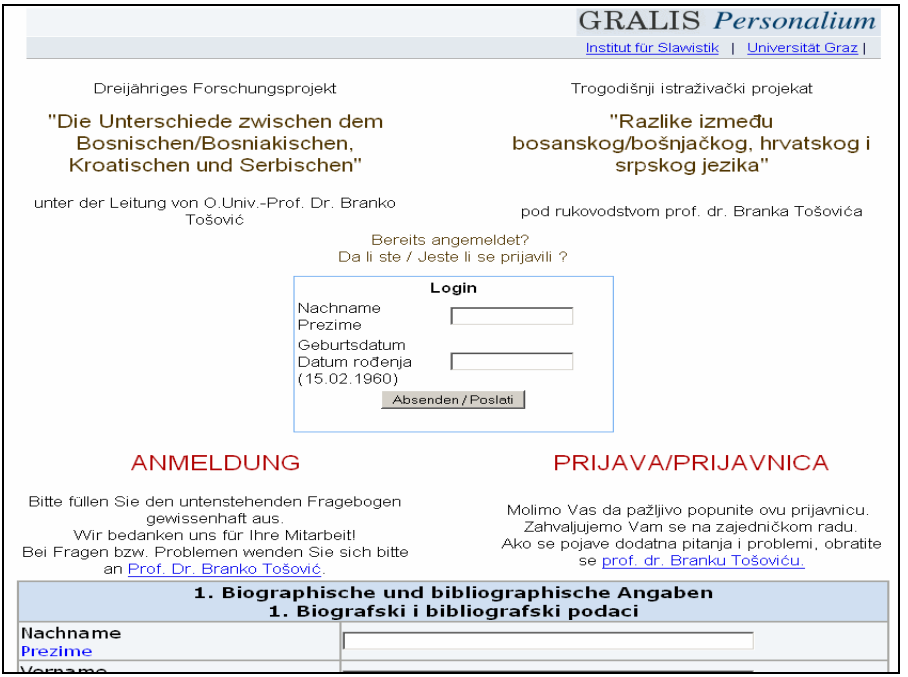


Abb. 90: Einstiegsfenster in das Gralis-Personalium – erster Teil des Fragebogens

Der in blauer Farbe gehaltene Fragebogen setzt sich aus drei großen Teilen zusammen, wobei der erste (in Abb. 1 ersichtlich) biographische und bibliographische Angaben beinhaltet. Diese umfassen u. a. persönliche Daten, Informationen zum Arbeitsplatz, zu eventuell vorhandenen Lehrveranstaltungsschwerpunkten (für Lehrende), zu wissenschaftlichen Forschungsgebieten und zu Publikationen. Es sei darauf hingewiesen, dass sämtliche Felder optional sind und etwa bei Nichtzutreffen der Fragestellung oder dem Wunsch nach Nichtpreisgabe bestimmter Informationen nicht ausgefüllt werden müssen. Der zweite Teil des Fragebogens trägt den Titel Teilnahme am Projekt „Die Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen“ und ist konkret auf die Anforderungen des Projektes zugeschnitten. Die innere Unterteilung dieses zweiten Kapitels erfolgt nach den Projektjahren (1., 2. und 3. Forschungsjahr) und richtet an Benutzende die Frage, welche (linguistischen aber auch andere) Themen sie in welchem Jahr zu bearbeiten und auf den alljährlichen Symposien vorzutragen wünschen. Vorraussetzung dafür ist, dass in dem im gelb unterlegten Feld eingetragenen Auswahlkästchen für das jeweilige Jahr „ja“ aktiviert ist (siehe Abb. 91 unter „Themen aus dem 1. Forschungsjahr“).

2. Teilnahme am Projekt "Die Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen" 2. Učešće/sudjelovanje na projektu "Razlike između bosanskog/bošnjačkog, hrvatskog i srpskog jezika"	
Themen aus dem 1. Forschungsjahr: Teme iz 1. istraživačke godine	○ ja da ◉ nein ne
A. Die phonetisch-phonologischen, orthoepischen und orthografischen Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen **A.** Fonetsko-fonološke, ortoepske i ortografske razlike između bosanskog/bošnjačkog, hrvatskog i srpskog jezika	☐ Phonetik Fonetika ☐ Phonologie Fonologija ☐ Phonostilistik Fonostilistika ☐ andere drugo ☐ Akzentologie Akcentologija ☐ Orthoepie Ortoepija ☐ Orthographie Ortografija
B. Korpuslinguistik **B.** Korpusna lingvistika	☐ Korpuslinguistik Korpusna lingvistika
C. BKS-Voice	☐ BKS-Voice
Ihr Thema Vaša tema	
2. Jahr: 2. godina	○ ja da ◉ nein ne
Die lexikalischen, idiomatischen und derivativen Unterschiede zwischen dem Bosnischen/Bosniakischen,	☐ Lexikologie Leksikologija ☐ Lexikographie Leksikografija ☐ Parömiologie Paremiologija ☐ Semantik Semantika

Abb. 91: Zweiter, projektbezogener Teil des Fragebogens

Ab dem „2. Jahr“, d. h. dem Projektjahr, in dem die neu gestaltete Anmeldung als Gralis-Personalium in Betrieb ging, bietet sich den mitarbeitenden Personen eine zusätzliche Option, die es ermöglicht, neben dem gewünschten Thema auch die Zusammenfassung des Referates für das jeweilige Symposium einzutragen. Weiters werden alle ReferentInnen ersucht, vor Veranstaltungsbeginn – sofern vorhanden – auch ihre Präsentation, ihr Handout und – falls bereits fertig gestellt – den Text ihres Beitrages in ihren Fragebogen einzufügen. Auf diese Weise werden diese Dokumente an einem zentralen Ort abgelegt, wodurch sich die Administration und die organisatorischen Aktivitäten (etwa bei der Tagungs-Programmgestaltung) wesentlich vereinfachen.

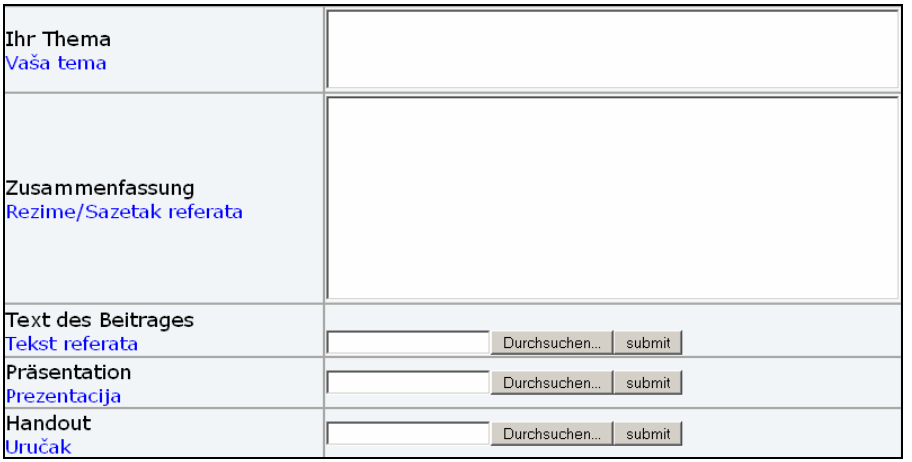


Abb. 92: Fenster für die Eingabe von Thema, Zusammenfassung und Text eines Beitrages sowie für Präsentation und Handout

Den abschließenden dritten Teil des Fragebogens bilden schließlich einige weitere Felder mit allgemeinen Fragen zur gewünschten Projektmitarbeit, darunter etwa zum Datum der geplanten Fertigstellung einer wissenschaftlichen Arbeit (siehe Abb. 93: „Habilitation“, „Doktorarbeit“, „Diplomarbeit“), zu diversen Tätigkeiten, die Benutzende im Rahmen des Projektes auszuüben wünschen. Auch hier ist wiederum eine Gliederung nach Jahren möglich, wodurch eine konkrete Evidenz über den geplanten wissenschaftlichen Fortgang, die Art der Mitarbeit und schließlich auch über die finanziellen Vorstellungen („Honorar pro Stunde“) entsteht.

Habilitation Doktorski rad/habilitacija	☐ 1. Jahr 1. godina ☐ 2. Jahr 2. godina ☐ 3. Jahr 3. godina Thema Tema
Doktorarbeit Magistarski rad	☐ 1. Jahr 1. godina ☐ 2. Jahr 2. godina ☐ 3. Jahr 3. godina Thema Tema
Diplomarbeit Diplomski rad	☐ 1. Jahr 1. godina ☐ 2. Jahr 2. godina ☐ 3. Jahr 3. godina Thema Tema
Korpusarbeit in Graz Rad na korpusu	☐ 1. Jahr 1. godina ☐ 2. Jahr 2. godina ☐ 3. Jahr 3. godina
Feldarbeit Rad na terenu / terenski rad	☐ 1. Jahr 1. godina ☐ 2. Jahr 2. godina ☐ 3. Jahr 3. godina Honorar pro Stunde (Euro) € Honorar po satu/času Honorar pro Seite (Euro) € Honorar po stani
Datenbankarbeit rad na bazi podataka	☐ 1. Jahr 1. godina ☐ 2. Jahr 2. godina ☐ 3. Jahr 3. godina Honorar pro Stunde (Euro) € Honorar po satu/času
Computer-Konkordanzen Kompjuterska/kompjutorska konkordanca	☐ 1. Jahr 1. godina ☐ 2. Jahr 2. godina ☐ 3. Jahr 3. godina Honorar pro Stunde (Euro) € Honorar po satu/času
Webseite des Projekts Internet-strana / mrežna strana projekta	☐ 1. Jahr 1. godina ☐ 2. Jahr 2. godina ☐ 3. Jahr 3. godina Honorar pro Stunde (Euro) € Honorar po satu/času
Administration, Korrespondenz, usw. Administrativni poslovi,	☐ 1. Jahr 1. godina ☐ 2. Jahr 2. godina ☐ 3. Jahr 3. godina Honorar pro Stunde (Euro) € Honorar po satu/času

Abb. 93: Allgemeine Angaben zur Projektmitarbeit

Nach Ausfüllen dieses dritten Teils verbleibt am Ende des Fragebogens noch eine letzte, kurze Rubrik, in der die Frage nach der Bereitschaft des Begutachtens von Texten gestellt wird. Bei diesen Texten handelt es sich in der Mehrzahl um die Beiträge der wissenschaftlichen Symposien, die alljährlich in einem Sammelband veröffentlicht werden, doch können daneben auch unterschiedlichste Bewertungen für anderenorts publizierte Texte abgegeben werden. Dieser letzte Eintrag im Rahmen des Fragebogens im Gralis-Personalium, bei dem auch andere GutachterInnen empfohlen werden können, steht in direktem Zusammenhang mit dem Programm Gralis-Rezensarium[1], das eine Online- und ebenfalls zentral verwaltete Begutachtung von Texten ermöglicht.

[1] Siehe dazu den gleichnamigen Beitrag von Stefan Kofler und Arno Wonisch in diesem Kapitel.

Abb. 94: Vorschläge zur Begutachtung von Texten und Absenden des Fragebogens

Nach diesem letzten Arbeitsschritt verbleibt schließlich nur noch der Befehl zur erfolgreichen Übermittlung der Einträge, wozu am Ende des Fragebogens der grau unterlegte Button „Absenden/Poslati“ zu aktivieren ist, mit dem alle Angaben zentral auf einem Server gespeichert werden und für Benutzende zwecks Bearbeitung jederzeit zur Verfügung stehen. Sollte der Wunsch bestehen, Änderungen oder Ergänzungen vorzunehmen, sind bei jedem erneuten Einloggen in das in Abb. 1 dargestellte Loginkästchen als Username der eigene Nachname (mit Schreibung von eventuell vorhandenen Umlauten oder diakritischen Zeichen) und als Passwort das Geburtsdatum (mit der Schreibung von „0“ und Trennpunkten zwischen Tag, Monat und Jahr) einzugeben. Mit einem Klick auf „Formular bearbeiten – Formular preraditi“ (siehe Abb. 95) können sodann die Einträge im eigenen Fragebogen verändert werden.

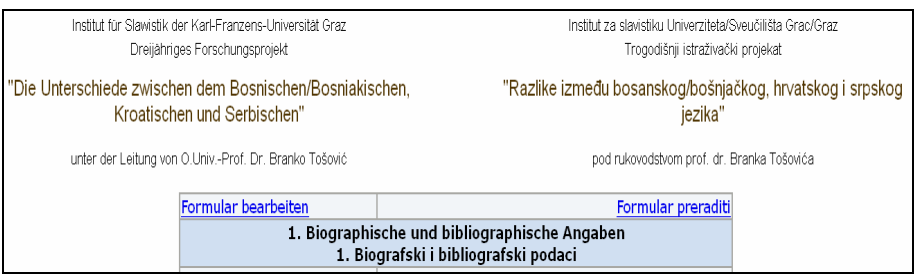
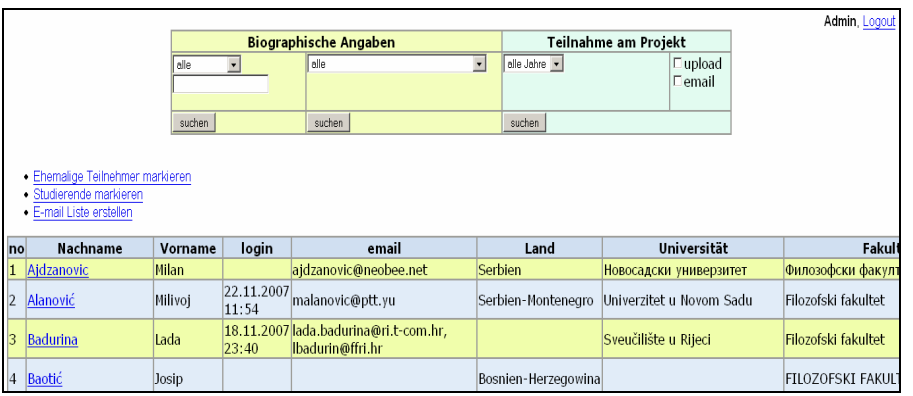


Abb. 95: Möglichkeit der Änderung von Einträgen ab dem zweiten Einloggen

Die Administration des Gralis-Personaliums. Neben der eben beschriebenen BenutzerInnen-Bearbeitung verfügt das Gralis-Personalium auch über eine Administrationsebene, die eine umfassende Darstellung und Auswertung sämtlicher BenutzerInneneinträge ermöglicht. Diese Administrationsoberfläche zeichnet sich durch eine klar unterscheidbare, andere graphische Gestaltung aus und bietet zahlreiche Such- und Bearbeitungsparameter. So sind z. B. Abfragen nach

folgenden Kriterien möglich: 1) zu biographischen Angaben: Nachname, Vorname, Arbeitsplatz, ehemalige (Mitarbeitende), (am Projekt mitarbeitende) Studierende, Land, Akademischer (wissenschaftlicher) Grad, Lehrtätigkeit, Linguistische Schwerpunkte, Nichtlinguistische Schwerpunkte, Sprache der Lehrtätigkeit, Untersuchte Sprachen, Fremdsprachen(kenntnisse); 2) zur Teilnahme am Projekt: alle Jahre, 1. Jahr, 2. Jahr, 3. Jahr.

Die in unten stehender Abb. 96 auf der linken Seite dargestellten drei Befehle „Ehemalige Teilnehmer markieren“, „Studierende markieren“ und „E-Mail Liste erstellen“ dienen dazu, mittels Auswahl der entsprechenden Personen die in den Menüpunkten definierten Arbeitsschritte in kürzester Zeit durchführen zu können und die Datenbank nach den genannten Kriterien neu zu generieren, wobei ein „Markieren von Ehemaligen oder Studierenden“ jederzeit wieder rückgängig gemacht werden kann. Durch einen Klick auf den in blauer Farbe dargestellten Familiennamen einer am Projekt mitarbeitenden Person werden sämtliche Angaben zu dieser einsehbar. Um diese Angaben schließlich ergänzen bzw. um zusätzliche Applikationen erweitern zu können, bietet das Programm im Menüpunkt „Teilnahme am Projekt“ (Abb. 96, türkises Kästchen rechts oben) die Option „upload“.



no	Nachname	Vorname	login	email	Land	Universität	Fakul
1	Ajdzanovic	Milan		ajdzanovic@neobee.net	Serbien	Новосадски универзитет	Филозофски факул
2	Alanović	Milivoj	22.11.2007 11:54	malanovic@ptt.yu	Serbien-Montenegro	Univerzitet u Novom Sadu	Filozofski fakultet
3	Badurina	Lada	18.11.2007 23:40	lada.badurina@ri.t-com.hr, lbadurin@ffri.hr		Sveučilište u Rijeci	Filozofski fakultet
4	Baotić	Josip			Bosnien-Herzegowina		FILOZOFSKI FAKUL

Abb. 96: Startseite der Administrationsebene

Wird diese aktiviert, ändert sich die Oberflächengestaltung der Administrationsoberfläche dahingehend, dass zu jeder Person Felder mit den (vorerst in BKS verfassten) Bezeichnungen „Referat“, „Diskusija“, „Prezentacija“ und „Handout“ hinzugefügt werden. Diese geben dem Administrationsteam die Möglichkeit, Aufsätze, Diskussionen, Präsentationen und Handouts für die jeweiligen Teilnehmenden zur Verfügung zu stellen. Dies bedeutet konkret, dass etwa im Rahmen einer Konferenz oder eines Symposiums gezeigte Präsentationen, verteilte Handouts, mündlich vorgetragene Diskussionsbeiträge und auch für eine Publikation vorgesehene Beiträge von den betreffenden Personen aufgerufen, bearbeitet und wieder in die Datenbank retourniert werden können. Zu den Diskussionsbeiträgen sei angemerkt, dass diese beim 1. Projektsymposium „Die Unterschiede zwischen dem Bosnischen, Bosniaki-

schen, Kroatischen und Serbischen“ (Graz, 12.–14. April 2007) nach vorheriger diesbezüglicher Information aufgezeichnet wurden und von den DiskutantInnen jederzeit in den Formaten „mp3“ oder „wav“ angehört werden können.

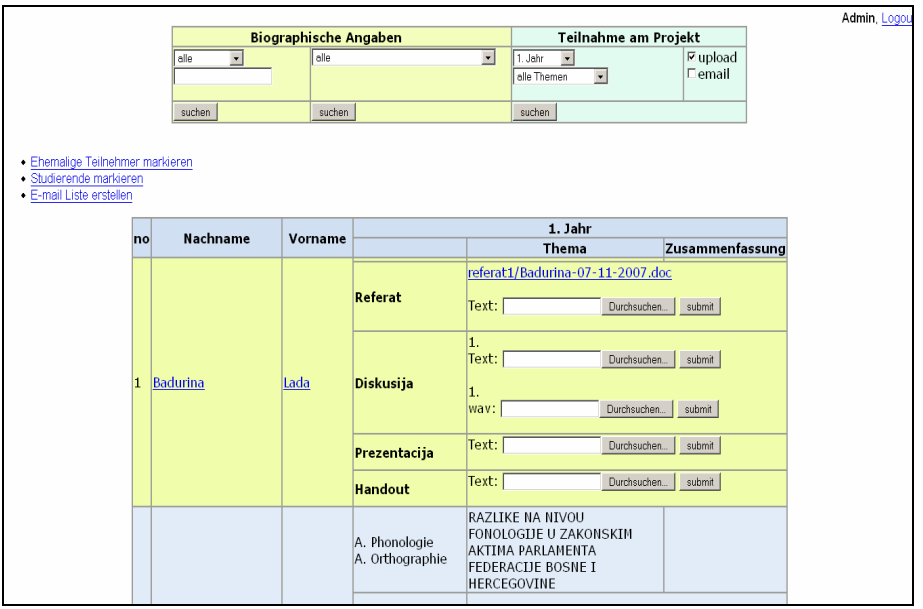


Abb. 97: Hinzufügen neuer Applikationen durch die Projektadministration

Abschließend kann festgehalten werden, dass mit der Entwicklung des Gralis-Personaliums ein wesentlicher Schritt in Richtung strukturierter, zentraler und kompakter Verwaltung personenbezogener Daten gesetzt wurde, wobei sowohl BenutzerInnen als auch AdministratorInnen ständig abrufbare und bearbeitbare Evidenzen zur Verfügung stehen. Auf diese Weise können, nicht zuletzt auch dank der zahlreichen Implikationsmöglichkeiten und Querverbindungen zu anderen Softwareprogrammen des Projektes, die Planung, Organisation und Durchführung von Veranstaltungen wie auch Editionsvorhaben leichter und effizienter in die Tat umgesetzt werden.

Branko Tošović (Graz)

Das Gralis-Präskriptarium

22. Das Gralis-Präskriptarium bildet einen Teil des Gralis-Komplementariums und dient zur Untersuchung von Rechtschreibungen slawischer Sprachen Mit der Entwicklung dieses Programms soll ein komplexes und rationelles Medium für eine Online-Suche nach orthographischen Informationen vor allem nahe verwandter slawischer Sprachen wie z. B. bosnisch/bosniakisch, kroatisch und serbisch geschaffen werden. Es sei darauf hingewiesen, die orthographischen Normen des BKS trotz der gleichen orthographischen Prinzipien (phonetisch-phonologisch) in einer Reihe von Fällen überaus unterschiedliche Lösungen anbieten, sodass ein Zurechtfinden in dieser Problematik nicht immer einfach ist. Dies trifft in besonderem Maße auf Universitäten zu, an denen wie etwa in Graz alle drei Sprachen unterrichtet werden, weshalb Studierende zumindest mit den grundlegenden Ähnlichkeiten und Unterschieden zwischen den einzelnen Standards vertraut sein sollten. Die Thematik wird zusätzlich noch dadurch erschwert, dass für manche Sprachen (etwa kroatisch und serbisch) mehrere Regelwerke vorhanden sind, die gegenseitig um ihre Repräsentation in Schulen, in den Medien, in Verlagshäusern u. Ä. wetteifern. Es ist vorgesehen, dass im Gralis-Präskriptarium sämtliche orthographisch-normativen Publikationen erfasst werden, für deren diesbezügliche Heranziehung eine Einverständniserklärung seitens der HerausgeberInnen vorliegt. Dabei sei angemerkt, dass keine einzige in Buchform vorhandene Rechtschreibung in ihrer Gesamtheit dargestellt wird, sondern den BenutzerInnen bloß nach Schlagwörtern und thematischen Einheiten gegliederte Inhaltsauszüge zur Verfügung gestellt werden. Dies geschieht anhand viererlei Arten von Informationen: 1. zu orthographischen Regeln (z. B. zur Groß- und Kleinschreibung), 2. zu Interpunktionszeichen (z. B. Punkt, Strichpunkt, Beistrich oder Anführungszeichen), 3. zur Schreibung gewisser „umstrittener“ Lexeme (z. B. einige Ijekavismen) und 4. Zur Schreibung einzelner Grapheme (z. B. **č** und **ć**). Die gesamte Information wird in der Originalgestalt in lateinischem und kyrillischem Alphabet dargebracht. Bei der Suche gilt es zuerst die Sprache und sodann das einzusehen gewünschte orthographische Regelwerk auszuwählen. Eine zweite Suchmöglichkeit sieht eine Suche in zwei oder mehreren Sprachen wie auch in allen im Gralis-Präskriptarium vorhandenen Sprachen vor Die erhaltenen Informationen bieten sodann Erklärungen, Beispiele für einzelne orthographische Lösungen und einen Verweis auf die Quelle. Das graphische Interface wurde für die Studienrichtungssprachen am Institut für Slawistik der Karl-Franzens-Universität Graz (BKS, russisch und slowenisch) und für das Deutsche entwickelt. In einer ersten Phase steht das Interface allerdings nur für BKS und deutsch zur Verfügung.

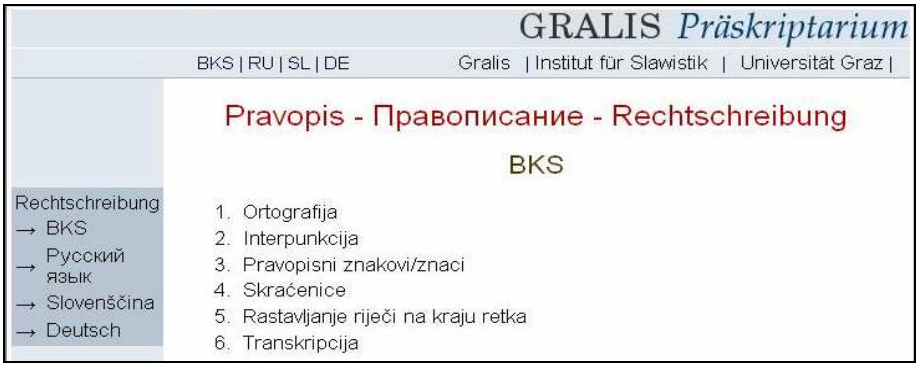


Abb. 98: Startseite des Gralis-Präskriptariums

Inhaltlich besteht das Gralis-Präskriptarium aus mehreren Teilen: a) aus einer Liste von orthographischen und Interpunktionszeichen, b) aus Regeln für die Orthographie, c) aus Regeln für die Interpunktion, d) aus Regeln für beide der zuletzt genannten Kategorien, d) aus Regeln für die Schreibung von Abkürzungen, e) für Abteilungen am Ende einer Zeile und f) aus Regeln für die Transkription.

Im Teil zu den orthographischen und Interpunktionszeichen werden die Regeln für die elementaren Zeichen genannt, nämlich jene zur Schreibung der Akzente:◌́, ◌̀, ◌̂, ◌̏, ◌̄, des Apostrophs: **’**, Gedankenstrichs: **–**, Bindestrichs: **-**, Doppelpunktes: **:**, des Zeichens für den Genitiv Plural◌̂, Schrägstriches: /, von Anführungszeichen: **„“**, **“”**, **« »**, **» «**, Halbanführungszeichen: **‘’**, **‘ ’**, **, ‘**, Buchstaben in der Funktion von Interpunktionszeichen, z. B.: **a**, **x**, **y**…, des Punktes: **.**, Strichpunktes: **;**, von drei Punkten, des Fragezeichens: **?**, Rufzeichens: **!**, von Klammern: ( ), / /, [ ], { }, des Gleichheitszeichens: **=**, Herkunftszeichens: **< >**, des Sternchens: *****, von Chat- und E-Mail-Symbolen: ☺, 😐, ☹, @ u. a.

Die orthographischen Regeln für das BKS betreffen die Schreibung einzelner Buchstaben und Wörter, wobei sich die Struktur wie folgt darstellt: 1. Schreibung (a) von Groß- und Kleinbuchstaben, b) von ersten Wörtern in einem Satz, (c) von Anredewörtern, 2. der Jat-Reflex betreffend: a) die ekavische Aussprache, b) die ijekavische Aussprache, 3. das Graphem **j**, 4. die Grapheme **č** und **ć**, 5. die Grapheme **đ** und **dž**, 6. die erste Palatalisierung, 7. die zweite Palatalisierung/Sibilarisierung, 8. Den Umlaut, 9. die Vokalisierung von **l** zu **o**, 10. die Assimilation von Konsonanten nach: a) Stimmhaftig-/Stimmlosigkeit, b) Artikulationsstelle, c) Artikulationsart, 11. der Schwund von Konsonanten, 12. Getrennt- und Zusammenschreibung von: a) Substantiven, b) Adjektiven, c) Numeralia, d) Verben, e) Adverbien, f) Modalwörtern, g) Präpositionen, h) Konjunktionen, i) Partikeln.

Die Interpunktionsregeln beziehen sich auf die syntaktische Struktur, genauer gesagt auf die Verwendung einzelner Zeichen innerhalb eines Satzes. Hierbei wird besonderes Augenmerk auf den Beistrich und dessen Position (im

Folgenden ausgeführt) gelegt: a) zwischen Satzteilen, b) zwischen Hauptsätzen (Disjunktivsätze, Adversativsätze), Nebensätzen (Exklamativsätze, Konditionalsätze, Konzessivsätze, Finalsätze, Konsekutivsätze, Imperativsätze, Relativsätze, Temporalsätze, Lokalsätze, Modalsätze, Komparativsätze) und c) in Sätzen mit Erweiterungen. Einen wichtigen Aspekt stellt dabei die Inversion dar.

Im System der Abkürzungen werden die grundlegenden Regeln mit Beispielen angeführt, auf das der Teil mit der Wortabteilung am Zeilenende folgt.

Die Transkription wird nach Sprachen gegliedert: **(a)** klassische Sprachen (altgriechisch, Latein, althebräisch), **(b)** slawische Sprachen – südslawische (BKS, bosnisch/bosniakisch, kroatisch, serbisch, bulgarisch, mazedonisch, slowenisch), ostslawische (russisch, ukrainisch, weißrussisch), westslawische (tschechisch, sorbisch, polnisch, slowakisch), **(c)** germanische Sprachen (dänisch, deutsch, englisch, flämisch, irisch, isländisch, niederländisch, norwegisch, schwedisch), **(d)** romanische Sprachen (französisch, italienisch, rätoromanisch, rumänisch, spanisch), **(e)** andere europäische Sprachen (albanisch, finnisch, ungarisch), **(f)** südamerikanische Sprachen, **(g)** asiatische Sprachen (abchasisch, aserbaidschanisch, tschetschenisch, georgisch, Hindi, japanisch, armenisch, kasachisch, chinesisch, koreanisch, tadschikisch, usbekisch, vietnamesisch u. a.), **(h)** afrikanische Sprachen (Suaheli u. a.).

Im Zuge der Ausarbeitung des Gralis-Präskriptariums wird das hier vorgestellte Modell präzisiert, modifiziert und erweitert werden.

Literatur und Quellen

Gralis: http://www-gewi.kfunigraz.ac.at/gralis/

http://korpus.juls.savba.sk/index.sk.html

http://korpus.pl/

http://msdn.microsoft.com/

http://project-x.sourceforge.net/

http://riznica.ihjj.hr/

http://ruscorpora.ru/index.html

http://sourceforge.net/

http://torvald.hit.uib.no/talem/jana

http://ucnk.ff.cuni.cz/

http://www.autoitscript.com/

http://www.fidaplus.net/

http://www.hnk.ffzg.hr/

http://www.iis.fraunhofer.de/

http://www.kgw.tu-berlin.de/

http://www.korpus.matf.bg.ac.yu/prezentacija/korpus.html

http://www.megaling.crimea.edu/program/Rychkova.htm

http://www.mpeg.org/

http://www.narusco.ru/

http://www.tekstlab.uio.no/Bosnian/Korpus2.html

http://www.vorbis.com/

http://www.wikipedia.org/

IMS Corpus Workbench (CWB): http://www.ims.uni-stuttgart.de/projekte/

Tošović 2002 – Tošović, B. *Funkcionalni stilovi*. Beograd: Beogradska knjiga.

Tošović 2003 – Tošović, B. *Ujak*. Beograd: Beogradska knjiga.

Hansen/Neumann 2005[9]: Hansen, H. R./Neumann, G. *Wirtschaftsinformatik 1 – Grundlagen und Anwendungen*. Stuttgart: Lucius & Lucius.

Branko Tošović

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. Gralis-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.