ÖSTERREICHISCHE AKADEMIE DER WISSENSCHAFTEN
PHILOSOPHISCH-HISTORISCHE KLASSE
SITZUNGSBERICHTE, 879. BAND

VERÖFFENTLICHUNGEN ZUR LINGUISTIK
UND KOMMUNIKATIONSFORSCHUNG

BAND 30

HERAUSGEGEBEN VON

WOLFGANG U. DRESSLER

CLAUDIA RESCH, WOLFGANG U. DRESSLER (HG.)

Digitale Methoden der Korpusforschung in Österreich

VERLAG DER
ÖSTERREICHISCHEN
AKADEMIE DER
WISSENSCHAFTEN

le Repräsentation der Texte in einem adaptiven Layout unterstützt zwei grundverschiedene Rezeptionshaltungen: das Lesen und das Suchen in den Texten. Beide Nutzungssituationen werden durch das innovative Web-Interface ermöglicht, indem es einerseits den Lesenden den Komfort einer digitalen Lektüreansicht bietet und andererseits den Suchenden Navigationsinstrumente zur Verfügung stellt, die einen direkten Zugriff (über Inhaltsverzeichnisse, Register und freie Suche) auf bestimmte Stellen im Text erlauben.

LinguistInnen können ihre Suche nach einzelnen Wortformen, nach Lemmata oder nach Wortarten ausrichten, wobei bei letztgenannten die im Register gelisteten Abkürzungen des Stuttgart-Tübingen-TagSets zu verwenden sind. Weiters besteht die Möglichkeit, eine kombinierte Suchanfrage von Lemma und Wortart durchzuführen: Etwa könnte man nach dem Wort „der" suchen und mit der Angabe der Wortart (ART oder PRELS) einschränken, ob man nach dem Artikel oder dem Relativpronomen sucht [lemma="der" and pos="ART"] bzw. [lemma="der" and pos="PRELS"]. Auch die kombinierte Suche nach einem Wortteil innerhalb einer Wortart ist zulässig: So könnte man nach Nomen endend auf „-heit" suchen [lemma="*heit" and pos="NN"] oder nach Adjektiven (attributiv oder prädikativ / adverbial) mit dem Präfix „aller-", das den Superlativ verstärkt [lemma="aller*" and pos="ADJA/ADJD"] – Beispiele aus dem Korpus sind *allerliebst*, *allerschönst* oder *allerletzt*.

Da die Sprachwissenschaft sich für andere Fragen interessiert als die Literaturwissenschaft oder textbasierte, historisch ausgerichtete Wissenschaften (Theologie, Kunstgeschichte) und sich nicht alle Fragestellungen antizipieren lassen, war den Herausgeberinnen daran gelegen, die Originaldokumente möglichst authentisch und quellennah abzubilden. In den Annotationen ist sorgfältig geprüftes Wissen über die Texte kodiert, das jederzeit online abgerufen werden kann. Mit der ABaC:us-Edition möchte das Projektteam – einem abrahamischen Buchtitel gemäß – „Etwas für alle" (1699) bieten. Es bleibt zu wünschen, dass ABaC:us in Forschung, Lehre und Unterricht rezipiert und erprobt wird und darüber hinaus auf die Neugier einer interessierten, web-affinen (Fach-)Öffentlichkeit trifft, der Abraham a Sancta Clara heute noch ein Begriff ist und die seine Zitate über den Tod nicht in Blütenlesen oder Anekdoten sucht, sondern sie im unveränderten Wortlaut unter Angabe der jeweiligen Belegstelle(n) finden möchte.

IV. DIE MORPHOLOGISCHE ANNOTATION IM GRALIS-KORPUS

Branko Tošović[1]

1. Einführung

An der Karl-Franzens-Universität Graz wurde ein komplexes und mehrsprachiges Korpus der geschriebenen und gesprochenen Sprache entwickelt, das für Analysen und das Erlernen aller slawischen Sprachen dient.[2] Die Arbeit an der Entwicklung dieses Korpus wurde im Jahre 2005 begonnen, seit 2007 wird es von SlawistInnen in mehreren Ländern sowohl in kollektiven und individuellen Forschungsprojekten als auch im Unterricht genutzt. Dieses Korpus mit der Bezeichnung Gralis-Korpus wurde dahingehend konzipiert, dass es als ein-, aber auch als mehrsprachiges (paralleles) Korpus eingesetzt werden kann, wobei die Hauptorientierung in einer Parallelisierung von mindestens zwei genetisch eng verwandten (slawischen) oder weniger eng verwandten Sprachen (slawisch-deutsch) liegt. BenutzerInnen haben die Möglichkeit, das Inter-

[1] Institut für Slawistik, Karl-Franzens-Universität Graz

[2] Vgl. Branko Tošović: Das Gralis-Korpus. In: Branko Tošović und Arno Wonisch (Hrsg.): Srpski pogledi na odnose između srpskog, hrvatskog i bošnjačkog jezika. Bd. I/2. Graz, Beograd 2010, S. 491-519.
Vgl. Branko Tošović: Das Gralis-Korpus. In: Branko Tošović (Hrsg.): Die Unterschiede zwischen dem Bosnischen / Bosniakischen, Kroatischen und Serbischen. Graz 2008, S. 724-827.
Vgl. Branko Tošović: Гралис-Корпус. In: Ursula Doleschal und Imke Mendoza (Hrsg.): Wiener slawischer Almanach. Sonderband 83. München 2013, S. 89-111.
Vgl. Branko Tošović: Upute za korišćenje Ćopićevog Gralis-Korpusa. In: Branko Tošović (Hrsg.): Poetika, stilistika i lingvistika pripovijedanja Branka Ćopića / Poetik, Stilistik und Linguistik des Erzählens von Branko Ćopić. Graz, Banja Luka 2012, S. 369-374.
Vgl. Branko Tošović: Leksička distanca između bosanskog / bošnjačkog, hrvatskog i srpskog jezika u Gralis-Korpusu. In: Branko Tošović (Hrsg.): Die Unterschiede zwischen dem Bosnischen / Bosniakischen / Kroatischen und Serbischen: Lexik-Wortbildung-Phraseologie. Wien, Berlin 2009, S. 17-63.
Vgl. Branko Tošović und Arno Wonisch: Gralis-Korpus. In: Jagoda Granić (Hrsg.): Jezična politika i jezična stvarnost. Zagreb 2009, S. 117-125.

face in mehreren Sprachen – Deutsch, Englisch, Russisch und Serbisch / Kroatisch / Bosni(aki)sch / Montenegrinisch – zu wählen (die Zahl der Sprachen wird sich in Zukunft erhöhen).

Das Korpus besteht aus zwei Subsystemen – einem auditiven und einem textuellen. Das Gralis Speech-Korpus umfasst transkribierte Audioaufnahmen, die die Möglichkeit einer phonetischen (akustischen, artikulatorischen), prosodischen und phonologischen Analyse von einzelnen Wörtern auf Laut-/Phonem-, Silben-, Lexem-, Syntagmen- und Satzebene bieten. Es besteht aus drei Subkorpora, dem Wort-, Fix- und Frei-Korpus. Das Wort-Korpus beinhaltet isoliert ausgesprochene Wörter in allen slawischen Sprachen (am meisten für das Serbische, Kroatische und Bosni/aki/sche). Das Fix-Korpus umfasst Aufnahmen eines vorgegebenen Textes und bietet bislang Material für die serbische, kroatische und bosni(aki)sche Sprache. Die phonetische und prosodische Transkription wird von ExpertInnen mithilfe des Programms Valorisarium durchgeführt. Das Frei-Korpus setzt sich aus Aufnahmen spontaner Rede zusammen (und besteht bislang aus einer rund 120-minütigen, mündlichen Erzählung). Mit dem Speech-Korpus ist das Programm Akzentarium[3] verbunden, das akzentuelle Informationen für mehr als 120.000 Wörter der Sprachen Serbisch, Kroatisch und Bosni(aki)sch bietet. Gegenstand der Forschungen im Zeitraum vom 1. Oktober 2012 bis zum 30. März 2013 war die automatische Generierung von Substantiv-, Adjektiv-, Pronominal- und Verbalparadigmen in den Sprachen Serbisch, Kroatisch, Bosni(aki)sch und Montenegrinisch mithilfe einer festgelegten Zahl an Flexionsregeln.

Das Gralis Text-Korpus enthält parallele Texte für Analysen zu allen slawischen Sprachen, wobei der Fokus bis dato auf der Befüllung mit Texten in südslawischen Sprachen (Serbisch, Kroatisch, Bosni(aki)sch, Montenegrinisch, Bulgarisch, Mazedonisch und Slowenisch) und in der slawischen Sprache mit den meisten SprecherInnen (Russisch) lag. Die entwickelte Infrastruktur bietet (1) die Wahl aller slawischen Sprachen und des Deutschen, (2) eine Parallelisierung slawischer Sprachen nach den drei Arealen (ost-, süd- und westslawisch) und (3) die Wahl von Sprachen aus einem der drei Großareale (z. B. südslawisch).

[3] Es handelt sich um ein Programm, mit dem es möglich ist, die Akzentuierung(en) für jedes Wort zu finden. Vgl. http://www-gewi.uni-graz.at/gralis-alt/php/en/Akzentarium/suche.php (30.6.2014).

Daneben gibt es Korpora zu einzelnen Schriftstellern, wobei gegenwärtig vier Subkorpora zur Verfügung stehen – zu Ivo Andrić (1892-1975; Nobelpreisträger für Literatur), zu Branko Ćopić (1915-1984; einer der größten slawischen Erzähler, Humoristen und Satiriker), zu Zoran Živković (geboren 1948; der meistübersetzte Literat des ehemaligen Jugoslawien) und zu Blaže Koneski (1921-1993; der bekannteste mazedonische Literat, Lyriker und Philologe).[4] Einen besonderen Korpustyp bildet das Edukativ-Korpus, in das Texte für das Verfassen von Diplomarbeiten, Dissertationen und Habilitationen aufgenommen werden. Das Gralis Text-Korpus besteht derzeit aus 5.300.000 Tokens und steht in all jenen Segmenten für eine Nutzung zur Verfügung, für die die Frage der AutorInnenrechte gelöst wurde.

Im Rahmen des Gralis-Korpus kann auch auf unterschiedliche Programme für Forschung und Lehre hingewiesen werden: Diese tragen die Namen Akzentarium (Online-Programm für das Erlernen des Akzentsystems von Serbisch, Kroatisch, Bosni(aki)sch), Anketarium (monatliche Online-Befragung der Studierenden des Instituts für Slawistik[5]), Bibliothekarium (bibliographisches Hilfsmittel zur Durchführung wissenschaftlicher Projekte und im Unterricht), MorphoGenerator (morphosyntaktische Annotation aller veränderlichen Wortarten, automatische Generierung aller Formen und Paradigmen mit Deklinationen, Konjugationen, Komparationen zur automatischen Analyse) und Lexikarium (dieses ist mit dem Text-Korpus, dem Akzentarium und dem MorphoGenerator verbunden und bietet komplexe prosodische, lexikalisch-semantische und grammatikalische Informationen).

Typologisch handelt es sich um ein lemmatisiertes Korpus (bislang sind eine morphosyntaktische Annotation und die Suche nach entsprechen-

[4] Vgl. Branko Tošović (Hrsg.): Поетиката, стилистиката и лингвистиката на текстовите од Блаже Конески во корпусот Гралис / Poetik, Stilistik und Linguistik der Texte von Blaže Koneski im Gralis-Korpus. Graz, Skopje 2013.

[5] Mithilfe des Programms Anketarium konnte die Möglichkeit geschaffen werden, Fragebögen, Umfragen und Datenerhebungen online durchzuführen, wobei es allen Studierenden mit individuell granulierbaren Zugangsberechtigungen offensteht, Fragebögen und andere Dokumente online zu erstellen und die auf einem Webserver gespeicherten Ergebnisse jederzeit abzurufen. Das Programm zeichnet sich dadurch aus, dass es allen an der Erstellung von Online-Fragebögen und Umfragen beteiligten Studierenden ermöglicht, diese problemlos, effizient und den eigenen Bedürfnissen entsprechend anzulegen.

den Annotationen für die Sprachen Serbisch, Kroatisch, Bosni(aki)sch und Montenegrinisch möglich). Die Generierung der Paradigmen für alle flektierenden Wortarten auf Basis der morphologischen Annotation verfolgt unterschiedliche Ziele: (1) Untersuchung grammatikalischer Formen und Konstruktionen, (2) Gebrauch des annotierten Materials für die Lehre der serbischen, kroatischen und bosni(aki)schen Sprache als Erst- und Fremdsprache, (3) automatisches Suchen nach Tokens (Wortformen) mit bestimmten morphologischen Merkmalen im Gralis-Korpus.

2. Das Gralis Text-Korpus

Im Gralis Text-Korpus gibt es drei Arten der Annotation: (1) eine metatextuelle, (2) eine extralinguistische und (3) eine linguistische[6], wobei die metatextuelle Annotation Informationen zu Titel, Kapitel und Absatz bietet. Für die morphologische Annotation und die Analyse der grammatikalischen Struktur der Korpus-Sprachen dient das Online-Programm Gralis-MorphoGenerator, das in Verknüpfung mit dem Korpus eine vollständige morphosyntaktische Annotation sämtlicher Tokens (Wortformen) des Gralis-Korpus ermöglicht. Auf diese Weise kann neben tabellarischen Übersichten in Bezug auf die Frequenz einzelner Wörter und Wortarten auch ein kompletter Überblick über die Flexion aller Tokens der drei Sprachen des Korpus gegeben werden, wodurch für Lernende eine wertvolle Hilfe zum Studium der in diesen Sprachen überaus komplexen Deklinations- und Konjugationsmuster geschaffen werden konnte.

Die extralinguistische Annotation verfügt über folgende Komponenten: (1) AutorInnen: individuelle AutorInnen (Vor- und Nachname), kollektive AutorInnen (Vor- und Nachname), fingierte AutorInnen (Vor- und Nachname), Pseudonym, unbekannte AutorInnen (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, HerausgeberInnen; Angaben zur Sprache, zur regionalen Variante, Schrift, Übersetzung (ÜbersetzerInnen); (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 u. a.), Chronotop (welche Zeit und welcher Ort werden im Text behandelt); (5) strukturelle Angaben: Art der Formatierung, Reim (falls vorhanden) und (6) kommunikatorische Angaben (Zielgruppe des Texts).

[6] Die linguistische Annotation umfasst die Hervorhebung von Sätzen, Syntagmen und Wörtern, wobei zwischen folgenden weiterführenden Annotationsschritten unterschieden wird: (1) morphologische Annotation: nach morphosyntaktischen Kategorien; (2) orthoepische Annotation: nach der Art des Akzents (lang steigend, lang fallend, kurz steigend, kurz fallend, Länge); (3) semantische Annotation: gemäß dem Programm WortNet; (4) stilistische Annotation: nach der Art des Stils, der Art des funktionalen Stils (literarisch-künstlerisch, publizistisch, wissenschaftlich, administrativ, umgangssprachlich) und (5) syntaktische Annotation: gemäß dem syntaktischen Baum der Abhängigkeiten. Siehe auch Branko Tošović: Морфологическое порождение существительных сербского, хорватского, бошняцкого и черногорского языков. In: Doleschal und Mendoza [Anm. 2], S. 113-134.

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

Angesichts dessen, dass die Qualität jedes Korpus durch (1) die Tiefe und den Umfang der Annotation, (2) die Suchmöglichkeiten, (3) die Repräsentativität, Proportionalität und Ausgewogenheit sowie (4) die Zugänglichkeit bestimmt wird, wird diesen Faktoren bei der stetigen Weiterentwicklung des Korpus umfassend Rechnung getragen.

Die morphosyntaktische Annotation und die Suchabfrage erfolgen mit dem Gralis-MorphoGenerator. Dieses Programm ermöglicht es, komplexe statistische Informationen zu jedem Token des Gralis-Korpus zu erhalten. Dieses online abrufbare analytisch-synthetische System dient für (1) die automatische morphosyntaktische Annotation von Wörtern, (2) grammatikalische Analysen und (3) die automatische Generierung der Paradigmen zu allen veränderlichen Wortarten.

Die grammatikalischen, lexikographischen, orthoepischen und orthographischen Informationen werden im Online-Wörterbuch Gralis-Le-

xikarium vereinigt.[7] Die diesbezüglichen Datenbanken bestehen aus 3.308.359 Wortformen, die auf 259.283 Lemmata (lexikalische Grundformen) zurückgehen. Die meisten Wörter gehören zu den Substantiven, gefolgt von Pronomina und Verben. Im Falle einer veränderlichen Wortart (Substantive, Adjektive, Pronomina, Verben, Zahlwörtern und teilweise Adverbien) besteht im Gralis-Lexikarium die Möglichkeit, das gesamte Paradigma eines Wortes abzurufen. Bedingt durch die automatische Generierung aller in den Datenbanken des Lexikariums vorhandenen Wörter und die jeweiligen Zusatzinformationen kann das Online-Wörterbuch Gralis-Lexikarium als wertvolles elektronisches Nachschlagewerk herangezogen werden, das seinen BenutzerInnen nicht nur Informationen zur Wortbedeutung in der jeweils anderen Sprache bietet, sondern dazu auch gesamte Paradigmen anzeigt, Angaben zu Synonymen und graphisch ähnlichen Lexemen tätigt, Häufigkeitsstatistiken errechnet und einige weitere Informationen zur Verfügung stellt.

3. Morphologische Annotation

Die Tätigkeiten zur automatischen Annotation von ca. 120.000 Wörtern des Bosni(aki)schen, Kroatischen und Serbischen wurden nach fünf Jahren (2008-2013) abgeschlossen, wobei die automatisch generierte Aufstellung der flektierten Formen aller Wortarten und ihrer Paradigmen mit den grundlegenden grammatikalischen Informationen in Gralis (das linguistische Slawistik-Portal der Karl-Franzens-Universität Graz)[8] zugänglich ist. Die morphologische Annotation wurde im April 2008 im Rahmen des Projektes „Die Unterschiede zwischen dem Bosnischen / Bosniakischen, Kroatischen und Serbischen“[9] begonnen und im Rahmen des Projekts Gralis-Lexikarium (2008-2013, gefördert durch die Steiermärkische Landesregierung) fortgesetzt. Das System der Annotation und

[7] Vgl. http://www-gewi.uni-graz.at/gralis-alt/0.Projektarium/MorphoGenerator/lex_login.php (30.6.2014).
Gegenwärtig wird an diesem Online-Wörterbuch gearbeitet.

[8] Als Plattform für dieses Korpus dient das Gralis-Portal: http://www-gewi.kfunigraz.ac.at/gralis/index.html (30.6.2014), siehe auch Branko Tošović: Gralis: Das linguistische Slawistik-Portal der Karl-Franzens-Universität Graz (2000-2010). Graz 2010.

[9] Branko Tošović: Die Unterschiede zwischen dem Bosnischen / Bosniakischen, Kroatischen und Serbischen (FWF-Projekt P19158-G03 2006-2010). Konzeption, Aktivitäten, Ergebnisse. Graz 2010.

die Generierung wurden von Branko Tošović entwickelt, die dazugehörigen Software-Applikationen stammen von Olga Lehner.

4. Kodierung

Für die morphologische Annotation wurde die Multext-East-Kodierung (Multilingual Texts and Corpora for Eastern and Central European Languages – multilingual dataset for language engineering research and development: MultiText East)[10] gewählt, die im Jahr 2004 von Tomaž Erjavec und seiner Gruppe entwickelt wurde. Sie umfasst Codes für alle Wortarten, Abkürzungen und so genannte Residuals[11]:

CATEGORY (en)	Value (en)	Code (en)	Attributes
CATEGORY	Noun	N	5
CATEGORY	Verb	V	10
CATEGORY	Adjective	A	7
CATEGORY	Pronoun	P	10
CATEGORY	Adverb	R	2
CATEGORY	Adposition	S	3
CATEGORY	Conjunction	C	4
CATEGORY	Numeral	M	6
CATEGORY	Particle	Q	1
CATEGORY	Interjection	I	1
CATEGORY	Abbreviation	Y	4
CATEGORY	Residual	X	0

Abbildung 1: Die grundlegenden grammatikalischen Kategorien für die Annotation im Gralis-Korpus

5. Positionen der Kodierung

Die morphologische Kodierung für das Gralis-Korpus besteht aus folgenden Positionen: (1) Wortart, (2) Typ der Wortart, (3) Verbalmodus, (4) Tempus, (5) Person, (6) Zahl, (7) Genus, (8) Diathese, (9) Responsiv, (10) Un/Bestimmtheit, (11) Reflexivität, (12) Kasus, (13) Un/Belebtheit, (14) Klitika, (15) Aspekt, (16) Etikettieren, (17) Akzent, (18) Aktionsart, (19) lexikalisch-semantische Gruppe, (20) Kollokation (Rektion, Kongruenz), (21) Expression, (22) funktionaler Stil, (23) analytische Form, (24) Destruktion, (25) Derivation, (26) Nummer der Regel, (27) grammatikalischer Typ der Generierung einer Form.

Auf Basis dieses Modells wurde ein konkretes Schema mit 20 Positionen entwickelt, das alle Wortarten und ihre Kategorien umfasst. Als Bezeichnung verwendet man in jeder Position Kleinbuchstaben (nur für

[10] Vgl. http://nl.ijs.si/ME (30.6.2014).

[11] Unter „Residuals“ versteht man den Rest als eine Menge, die nach einem Prozess, einem Geschehen oder einer Handlung übrig bleibt.

die Bezeichnung der Person und Regel dienen Ziffern). Da es mehr Positionen als Buchstaben gibt, wiederholen sich einige Grapheme, wobei die erste Position stets die Wortart nennt:

(1) Wortart
n – Substantive, **v** – Verben, **a** – Adjektive, **p** – Pronomina, **r** – Adverbien, **s** – Präpositionen, **c** – Konjunktionen, **m** – Numeralia, **i** – Interjektionen, **q** – Partikeln, **y** – Abkürzungen.

(2) Subtyp der Wortart
S u b s t a n t i v e: **c** – Gattungsbezeichnungen, **p** – Eigennamen, **m** – Stoffname, **l** – Kollektivum.
V e r b e n: **m** – Vollverben, **a** – Hilfsverben, **o** – Modalverben, **c** – Kopulaverben, **b** – Grundverben.
A d j e k t i v e: **f** – Relativadjektive, **r** – Relationsadjektive, **m** – Stoffadjektive, **s** – Possessivadjektive, **o** – Ordinaladjektive, **m** – Kardinaladjektive.
P r o n o m i n a: **p** – Personalpronomina, **d** – Demonstrativpronomina, **i** – Indefinitpronomina, **s** – Possessivpronomina, **q** – Interrogativpronomina, **r** – Relativpronomina, **x** – Reflexivpronomina, **z** – Negationspronomina, **g** – allgemeine Pronomina, **y** – interrogativ-relative Pronomina, **j** – bestimmte Pronomina, **t** – demonstrativ-relative Pronomina.
A d v e r b i e n: **g** – allgemeine, **z** – Negationsadverbien, **a** – adjektivische, **v** – verbale, **q** – Interrogativadverbien.
P r ä p o s i t i o n e n: **p** – präponierende, **t** – postponierende.
K o n j u n k t i o n e n: **c** – nebenordnende, **s** – unterordnende.
N u m e r a l i a: **c** – Grundzahlwörter, **o** – Ordnungszahlwörter, **m** – Kollektivzahlwörter, **l** – Multiplikativa, **s** – spezifische.
P a r t i k e l n: **z** – Negationspartikeln, **q** – interrogative Partikeln, **o** – Modalpartikeln, **r** – Bestätigungspartikeln.
A b k ü r z u n g e n: **n** – substantivische, **r** – adverbiale.

(3) Typ der Form
V e r b e n: **i** – Indikativ, **m** – Imperativ, **c** – Konjunktiv 1, **h** – Konjunktiv 2, **n** – Infinitiv, **p** – Partizip, **g** – Adverbialpartizip 1 (der Gegenwart), **w** – Adverbialpartizip 2 (der Vergangenheit), **u** – Supin, **t** – transitiv, **q** – zitierte, **s** – hypothetische.
A d j e k t i v e: Komparation: **p** – Positiv, **c** – Komparativ, **s** – Superlativ.

A d v e r b i e n: Komparation: **p** – Positiv, **c** – Komparativ, **s** – Superlativ, **e** – Elativ.

(4) Tempus: **p** – Präsens, **i** – Imperfekt, **f** – Futur I Sr (serbisch), **w** – Futur I Hr (kroatisch), **z** – Futur I Sr/Hr (serbisch und kroatisch), **q** – Futur II, **s** – Perfekt, **l** – Plusquamperfekt 1, **t** – Plusquamperfekt 2, **a** –Aorist

(5) Person: **1** – erste, **2** – zweite, **3** – dritte

(6) Numerus: **s** – Singular, **p** – Plural, **d** – Dual, **l** – Kollektivum

(7) Genus: **m** – maskulin, **f** – feminin, **n** – neutral, **l** – allgemein

(8) Diathese: **a** – Aktiv, **p** – Passiv

(9) Responsiv: **y** – ja, **n** – nein

(10) Bestimmtheit: **y** – ja, **n** – nein

(11) Reflexivität: **y** – ja, **n** – nein

(12) Kasus: **n** – Nominativ, **g** – Genitiv, **d** – Dativ, **a** – Akkusativ, **v** – Vokativ, **i** – Instrumental, **l** – Lokativ

(13) Belebtheit: y – ja, **n** – nein

(14) Klitika: **y** – ja, **n** – nein

(15) Aspekt: **p** – unvollendet, **e** – vollendet, **b** – doppelt

(16) Etikettierung: **y** – ja, **n** – nein

(17) Transitivität: **y** – ja, **n** – nein

(18) Destruktion: **y** – ja, **n** – nein

(19) Wortbildung: **s** – unmotivierte, **c** – Kompositum

(20) Zahl der Regel: 01, 02, 03 etc.

Der Beginn des Generierungsmodells der Paradigmen sieht folgendermaßen aus:

Substantiv	Art	Type	Person	Gender	Number	Case	Owner_ Number	Owner_ Gender	Clitic	Referent _Type	Sintactic _Type	Definiteness	Animate	Clitic_s	Pronoun _Form	Owner_ Person	Owner_ Number	Wh_ Type		Typ
	1	2	3	4	5	6	7	8	9	10	39	12	13	14	15	16	17	18	19	20
turist	N			m	s	n														39

Abbildung 2: Das substantivische Paradigma für das Wort *turist* ‚Tourist'

Nicht ausgefüllte Positionen sind dabei: 11. Reflexivität, 16. Etikettieren, 18. Destruktion, 19. Wortbildung. Von den 28 geplanten Positionen fehlen folgende: 18. Akzent (der sich im Akzentarium befindet), 19. Aktionsarten, 20. lexikalisch-semantische Gruppen, 21. Kollokation (Rektion, Kongruenz), 22. Expression, 23. funktionaler Stil und 24. analytische Form.

6. Der Gralis-MorphoGenerator

Beim Gralis-MorphoGenerator[12] handelt es sich, wie bereits besprochen, um ein Online-Tool, das eine umfassende morphosyntaktische Annotation für automatische Analysen von Wörtern und Paradigmen bietet. Das Ziel dieser Online-Applikation liegt darin, allen interessierten Personen automatisch alle Wortformen und -abwandlungen (Deklinationen, Konjugationen, Komparationen) darzulegen, sodass dieses Programm eine wesentliche Hilfe in Lehre und Forschung darstellt. Der MorphoGenerator ist mit dem Gralis BKS-Korpus verbunden und dient zur Annotation sämtlicher im Korpus enthaltenen Wortformen. Eine Suche kann entweder nach Wortformen, Lemmata oder einzelnen Wortarten durchgeführt werden. Dazu ist auch eine Abfrage nach den absoluten Häufigkeiten in den einzelnen Sprachen möglich. Als Ergebnis einer Suche erscheinen sämtliche Belege sowohl der Lemmata mit diesem Wortstamm als auch aus dem gesamten flektierten Paradigma. In weiterer Folge wird durch einen Klick auf ein blau unterlegtes Wort dessen gesamte Flexion abgebildet, wobei z. B. im Falle des Substantivs *ruka* (‚Hand') alle Kasus in Singular und Plural erscheinen. Auf der rechten Seite befinden sich die morphosyntaktischen Spezifikationen, nach denen die Annotation dieses Wortes vorgenommen wurde.

[12] Vgl. http://www-gewi.uni-graz.at/gralis-alt/0.Projektarium/MorphoGenerator/morpho.php (30.6.2014).

ruka			ruke		
Kasus	Singular		Kasus	Plural	
	f			f	
Nominativ	ruka	N-fsn----4SfA06	Nominativ	ruke	N-fpn----4SfA06
Genitiv	ruke	N-fsg----4SfA06	Genitiv	ruka ruku	N-fpg----4SfA06 N-fpg----4SfA06
Dativ	ruci	N-fsd----4SfA06	Dativ	rukama	N-fpd----4SfA06
Akkusativ	ruku	N-fsa----4SfA06	Akkusativ	ruke	N-fpa----4SfA06
Vokativ	rukom	N-fsv----4SfA06	Vokativ	ruke	N-fpv----4SfA06
Instrumental	rukom	N-fsi----4SfA06	Instrumental	rukama	N-fpi----4SfA06
Lokativ	ruci	N-fsl----4SfA06	Lokativ	rukama	N-fpl----4SfA06

Abbildung 3: Das Paradigma und der Code des Wortes *ruka* ‚Hand'

Ein Überblick über die morphosyntaktischen Spezifikationen im MorphoGenerator, der – wie erwähnt – mit dem Gralis-Korpus verbunden ist, in Bezug auf die einzelnen Wortarten und deren Häufigkeit, gegliedert nach Lemmata und Tokens, stellt sich wie folgt dar:

- Morphosyntaktische Spezifikationen
- Gralis Morpho-Generator
- Gralis BKS-Korpus

BKS morphologisches Lexikon			
Wortart		Lemma	Token
Abbreviatur	Y	243	243
Adjektiv	A	3576	613123
Adverb	R	6884	6901
Interjektion	I	589	595
Konjunktion	C	50	50
Partikel	Q	88	89
Pronomen	P	56	2506
Präposition	S	174	196
Substantiv	N	31837	464925
Verb	V	4858	277913
Zahlwort	M	1025	3816
insgesamt:		**49380**	**1370357**

BKS-Korpus Lexikon			
			Token
sr			101635
	hr		105521
		bs	52127
sr	hr	bs	31428
sr	hr		62165
	hr	bs	42735
sr		bs	35633
	insgesamt:		**150178**

Abbildung 4: Die Ergebnisse im Gralis-Lexikon

Im Falle eines Substantivs werden im Zuge der morphosyntaktischen Annotation die Kategorien Typ, Genus, Numerus, Kasus, Bestimmtheit, Vorkommen als Enklitikon, Belebtheit und personenbezogener Numerus unterschieden.

Substantiv									
Art	Type	Gender	Number	Case	Definiteness	Clitic	Animate	Owner_Number	Typ
1	2	3	4	5	6	7	8	9	10
N: Noun	c: common p: proper m: material l: collective	m: masculine f: feminine n: neuter l: collective	s: singular p: plural d: dual p: paucal	n: nominative g: genitive d: dative a: accusative v: vokative i: instrumental l: lokative	n: no y: yes	n: no y: yes	n: no y: yes	s: singular p: plural	01 02 ...

Abbildung 5: Die ersten Positionen der substantivischen Kategorien

Eng mit dem Gralis-MorphoGenerator in Verbindung steht ein Programm mit der Bezeichnung Gralis-PhonoGraphemator, das seine Datenbankbasis ebenfalls aus dem Gralis-Korpus bezieht. Dieses Programm ermöglicht Analysen zur Abfolge von Vokalen und Konsonanten in den Sprachen Serbisch, Kroatisch und Bosni(aki)sch, wie z. B. eine Suchabfrage nach der Abfolge des Vokals a und des Konsonanten c, wobei zuerst die Häufigkeit dieser Kombination im Gralis-Korpus und auch im Programm Gralis-Akzentarium angezeigt wird. Sodann werden durch einen Klick sämtliche Wörter angezeigt, die die Abfolge ac beinhalten.

Dank der morphologischen Annotation und der Funktionalität des Gralis-MorphoGenerators ist es nunmehr möglich, für jedes flektierbare Wort automatisch das gesamte Paradigma zu erhalten. Hierbei werden die Unterschiede zwischen dem Bosni(aki)schen, Kroatischen und Serbischen nicht global und pauschalisierend festgelegt, sondern in Form von Tabellenangaben über Vorkommen jedes einzelnen Tokens im Gralis-Korpus, wie etwa im Falle des Wortes *put* in den Bedeutungen 1. ‚Weg' (Substantiv, mask.), 2. ‚Hautfarbe' (Substantiv, fem.), 3. ‚mal' (Adverb) und 4. ‚nach, in Richtung' (Präposition).

Auf diese Übersicht über die morphosyntaktische Annotation und die grammatikalischen Kategorien dieses Verbs folgt sodann ein Überblick über alle vorhandenen Formen (z. B. im Falle eines Substantivs – sämtliche Flexionsformen in den einzelnen Kasus), was gerade für Lernende dieser Sprachen eine große Hilfestellung darstellt, da man einerseits mühelos gesamte Paradigmen für jedes veränderliche Wort erhalten kann und andererseits eine derartige Aufstellung des gesamten Formenbestandes von veränderlichen Wörtern bzw. Wortarten in keinem Lehr- oder Wörterbuch und auch in keiner Grammatik gefunden werden kann. Zwischensprachliche Unterschiede in der Häufigkeit des Gebrauches wer-

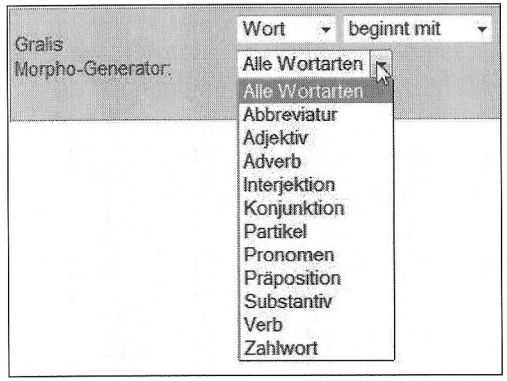

Abbildung 6: Auswahl der Parameter im MorphoGenerator

den nicht mit allgemeinen Formulierungen, sondern konkret und objektiv mit Korpusangaben dargestellt. Im MorphoGenerator kann man Tokens oder Lemmata, einen Wortan- oder einen Wortauslaut und / oder einen Teil eines Wortes suchen:

Der MorphoGenerator bietet auch eine rückläufige Sortierung und Angaben über die Häufigkeit in den Sprachen Serbisch, Kroatisch und Bosni(aki)sch. Für Substantive, Adjektive und Verben, die sich durch ein breites System an Änderungen (Deklination, Konjugation und Komparation) auszeichnen, wurden drei getrennte Masken entwickelt; siehe dazu etwa die Ansicht für Substantive:

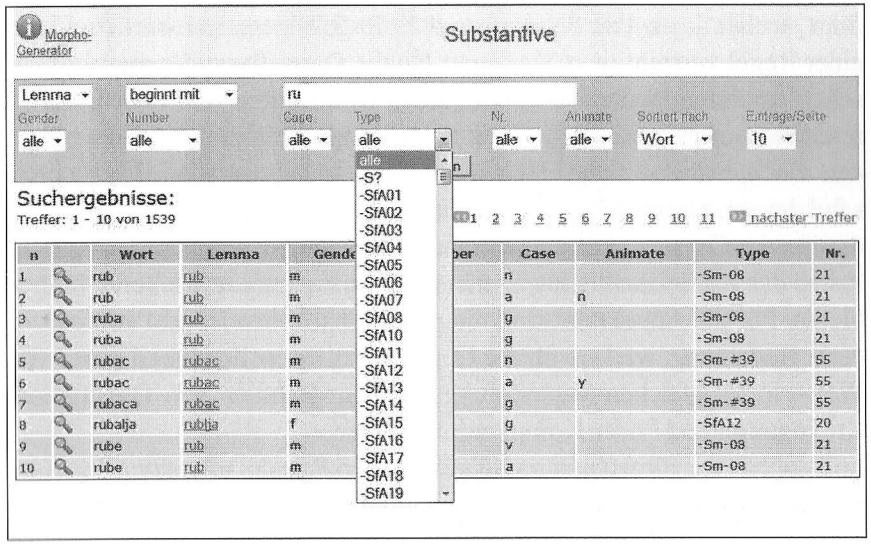

Abbildung 7: Suche nach Substantiven im MorphoGenerator

7. Statistik

Der MorphoGenerator umfasst auch statistische Informationen zu jeder Wortart im Gralis-Korpus.

SR-Korpus LEXICON				
Wortart		n	SUM Häufigkeit	%
Substantiv	N	22621	452597	31.99%
Pronomen	P	477	233359	16.49%
Verb	V	6982	156688	11.08%
Konjunktion	C	45	139478	9.86%
Präposition	S	118	129316	9.14%
Adjektiv	A	15900	126944	8.97%
Interjektion	I	179	118372	8.37%
Partikel	Q	62	89315	6.31%
Adverb	R	1923	88348	6.24%
Zahlwort	M	1011	21994	1.55%
Abbreviatur	Y	190	10587	0.75%
annotiert: 48.20 %		48991	1210688	85.58%
unannotiert: 51.80 %		52644	204058	13.58%
insgesamt:		101635	1414746	99.15%

HR-Korpus LEXICON				
Wortart		n	SUM Häufigkeit	%
Substantiv	N	22560	463071	32.39%
Pronomen	P	530	232995	16.29%
Verb	V	6731	160030	11.19%
Präposition	S	126	132322	9.25%
Konjunktion	C	45	129581	9.06%
Adjektiv	A	16652	128995	9.02%
Interjektion	I	180	120248	8.41%
Adverb	R	1992	86311	6.04%
Partikel	Q	62	75273	5.26%
Zahlwort	M	986	21865	1.53%
Abbreviatur	Y	194	10573	0.74%
annotiert: 46.94 %		49529	1212126	84.77%
unannotiert: 53.06 %		55992	217749	14.40%
insgesamt:		105521	1429875	99.17%

BS-Korpus LEXICON				
Wortart		n	SUM Häufigkeit	%
Substantiv	N	12249	151307	32.63%
Pronomen	P	453	75718	16.33%
Verb	V	4278	51166	11.03%
Adjektiv	A	8901	44909	9.68%
Konjunktion	C	39	44197	9.53%
Präposition	S	109	43264	9.33%
Interjektion	I	112	40555	8.75%
Adverb	R	1264	26295	5.67%
Partikel	Q	54	25757	5.55%
Zahlwort	M	636	8932	1.93%
Abbreviatur	Y	172	4823	1.04%
annotiert: 53.54 %		27908	396112	85.42%
unannotiert: 46.46 %		24219	67626	13.40%
insgesamt:		52127	463738	98.82%

Abbildung 8: Die allgemeine Statistik im Gralis-Korpus

8. Regeln

Die morphologische Annotation wurde für 100.461 Wörter durchgeführt, wobei deren Paradigmen mit 822 Regeln generiert wurden. Unter einer Regel versteht man Verfahren für die Darstellung der morphologischen Änderungen und Varianzen, deren Bestimmung und Beschreibung mithilfe kurzer Anweisungen bzw. die Matrize der Beziehungen zwischen den Codepositionen und der paradigmatischen Besetzung. Man unterscheidet syntagmatische und paradigmatische Regeln. Erstere umfassen die lineare Organisation der Wörter, wohingegen zweitere die Möglichkeiten der Wahl sprachlicher Einheiten in Bezug auf lineare Verkettungen bieten. Für die Annotation sind die paradigmatischen Regeln von besonderer Bedeutung, weil sie für die Generierung der vollen Paradigmen von Wörtern des gleichen Veränderungstyps (im Rahmen von Deklination, Konjugation und / oder Komparation) dienen.

Es gibt zwei Typen von paradigmatischen Regeln, einzelne und allgemeine: Einzelne Regeln beziehen sich auf bestimmte grammatikalische Kategorien und Unterkategorien und verweisen auf die Besonderheiten der Veränderung im Rahmen der bestimmten grammatikalischen Kategorien wie Kasus, Genus, Numerus etc. So etwa müssen bei der Erzeugung der Formen für männliche Substantive mit Nullendung vier Kasus im

Singular (Nominativ, Genitiv, Vokativ und Instrumental) und zwei Kasus im Plural (Nominativ und Genitiv) berücksichtigt werden.

Die allgemeinen Regeln umfassen drei Typen der Flexion: Deklination, Konjugation und Komparation. Den Regeln für die Deklination liegen Umstände zugrunde, die für die Generierung der Formen relevant sind. So etwa sind für die Erzeugung der substantivischen Paradigmen fünf von sieben Fällen im Singular (Nominativ, Genitiv, Akkusativ, Vokativ und Instrumental) und drei Fälle im Plural (Nominativ, Genitiv und Dativ) von Relevanz.

Es gibt drei Typen allgemeiner Regeln, nämlich strukturelle, kategoriale und interkategoriale. Die strukturellen Regeln verweisen auf formelle Mittel für die Generierung der Paradigmen und Alternationen an der Fugenstelle hin zum vorangegangenen Morphem. Dazu gehört die Regel in Bezug auf Endungen, denn so werden z. B. im System der Substantive unter Berücksichtigung des Suppletivismus alle Formen mit elf Endungen gebildet (čovjek – ljudi ‚Mensch – Leute'), mit Erweiterung des Stammes (*vuk – vukovi* ‚Wolf – Wölfe') und mit postakzentuierten Längen (vgl. Nom. Sg. und Gen. Pl. *Ovo je* žèna. ‚Das ist *eine Frau*'– *Nema* žénā ‚Es gibt *keine Frauen*'). Die kategorialen Regeln betreffen diejenigen Kategorien, die für die Generierung der morphologischen Formen von Bedeutung sind. Die intrakategorialen Regeln umfassen zwei oder mehrere Kategorien.

Die Notwendigkeit für die Bildung der großen Zahl der Regeln, besonders für Substantive und Verben, stellt ein Resultat der verschiedenen Alternationen dar, wobei zwei grundlegende Regeln unterschieden werden können: phonetisch-phonologische (in erster Linie das bewegliche /a/, die Palatalisierung, Sibilarisierung und Jotierung) und prosodische (vor allem die postakzentuierte Länge im Genitiv Plural einiger Substantive).

9. Annotationsphasen

Der Prozess des morphosyntaktischen Annotierens für das Gralis-Korpus besteht aus mehreren Verfahren. Die Annotationsphasen sehen wie folgt aus:

(1) Vorbereitung der Liste aller Wörter im Rahmen einer Standardsprache (Serbisch, Kroatisch, Bosni(aki)sch)

(2) Einteilung der Wörter nach Wortarten

(3) Vereinigung von Lexemen im Rahmen jeder Wortart in eine einzelne Liste entsprechend den grammatikalischen Merkmalen, die für die Generierung des Paradigmas wichtig sind (für Substantive die Kategorie des Genus, für Adjektive die Komparation, für Verben der Aspekt u. a.)

(4) Die dadurch erhaltene Liste wird unter Berücksichtigung zusätzlicher Merkmale (bei Substantiven die Belebtheit, bei Verben die Transitivität u. a.) weiter in Unterlisten unterteilt. Damit endet die Aufbereitung des lexikalischen Materials für die grammatikalische Verarbeitung.

(5) Sodann wird eine Analyse der Listen unter Punkt (4) durchgeführt und es kommt zur Bestimmung der paradigmatischen Marker (Kasus, Person u. a.), die für die Generierung der kompletten Paradigmen relevant sind (z. B. Nominativ, Genitiv, Vokativ und Instrumental Singular wie auch Nominativ und Genitiv Plural für männliche Substantive mit konsonantischer Endung).

(6) Nunmehr wird eine Liste mit allgemeinen Merkmalen erzeugt (z. B. für männliche Substantive mit Konsonantenendung, die im Nominativ Singular ein bewegliches /a/ aufweisen).

(7) Als weiterer Schritt wird ein typisches Wort in einer Gruppe gewählt und für dieses eine eigene Tabelle mit allen 20 Positionen und sämtlichen Formen entwickelt. Gemäß diesem Wort und der Nummer des Typs erhält jede Regel ihre eigene Benennung (z. B. $174\}^{\text{mudrac}}$). Der Anfang der Tabelle sieht folgendermaßen aus:

Branko Tošović — Substantiv_174_MUDRAC_Kod i paradigma

SUBSTANTIV

Substantiv

Substantiv	Art	Type	Person	Gender	Number	Case	Owner_Number	Owner_Gender	Clitic	Referent_Type	Sintactic_Type	Definitenes	Anima-te	Clitic_s	Pronoun_Form	Owner_Person	Owner_Number	Wh_Type		Tym
	1	2	3	4	5	6	7	8	9	10	174	12	13	14	15	16	17	18	19	20
mudrac	N			m	s	n														174
mudraca	N			m	s	g														174
mudracu	N			m	s	d														174
mudraca	N			m	s	a							y							174
mudrače	N			m	s	v														174
mudracem	N			m	s	i														174
mudracu	N			m	s	l														174
mudraci	N			m	p	n														174
mudraca	N			m	p	g														174
mudracima	N			m	p	d														174
mudrace	N			m	p	a														174
mudraci	N			m	p	v														174
mudracima	N			m	p	i														174
mudracima	N			m	p	l														174

Abbildung 9: Der Beginn des grammatikalischen Codes für die Regel 174^{mudrac}

Damit ist der linguistische Teil des Annotationsverfahrens beendet und der programmierend-technische beginnt.

(8) Das Material, das in den Arbeitsschritten (1) bis (7) erhoben und differenziert wurde, wird in Form von drei Tabellen in die relationale Datenbank MySQL überführt. Eine Tabelle umfasst vier Spalten mit dem Lemma, dem unveränderlichen Teil (Wortstamm), dem relevanten Merkmal (z. B. für Substantive die Belebtheit, für Verben die Transitivität) und der Nummer der linguistischen Regel, wie z. B.:

obnemoći	**obnemo**	n	01e
dići	**di**	y	03e
dolivati	**dol**	y	24

Tabelle 1: Auszug einer MySQL-Tabelle mit den Endungen

Der zweiten Tabelle liegt die Tabelle mit Code, Codepositionen und Paradigmen für eine bestimmte Regel zugrunde, die sich von der ersten Tabelle durch eine zusätzliche Spalte mit den unveränderlichen Segmenten des Wortes unterscheidet.

	1	2	3	4	5	6	7	8	9	10
dih	**V**	**m**	**i**	**a**	**1**	**s**	_	_	_	_
di	**V**	**m**	**i**	**a**	**2**	**s**	_	_	_	_
di	**V**	**m**	**i**	**a**	**3**	**s**	_	_	_	_

	11	12	13	14	15	16	17	18	19	20
dih	_	_	_	_	**e**	_	_	_	_	102e
di	_	_	_	_	**e**	_	_	_	_	102e
di	_	_	_	_	**e**	_	_	_	_	102e

Tabelle 2: Weitere MySQL-Tabelle

(9) Diese zwei Tabellen werden in das Format .csv überführt. In den neuen Tabellen trennt ein Strichpunkt die Inhalte ab:

šetati;še;y;45;

plesati;ple;y;48;

platiti;pla;y;90e;

In MySQL werden nunmehr Formen und Paradigmen erzeugt, wobei z. B. die MySQL-Tabelle zu den Verben 378 Regeln / Typen mit mehr als

lemma_id	lemma	stamm	transitive	typ
1	moći	mo	n	01
2	domoći	domo	y	01e
3	ispomoći	ispomo	y	01e
4	izmoći	izmo	y	01e
5	iznemoći	iznemo	n	01e
6	nasmoći	nasmo	y	01e
7	obnemoći	obnemo	n	01e
8	odmoći	odmo	y	01e
9	onemoći	onemo	n	01e

45.000 Reihen umfasst. Der Teil der Liste mit Lemmata und Wortstamm sieht folgendermaßen aus:

Abbildung 10: MySQL-Ansicht mit Lemmata, Endungen und dem Typ der Regel

Daraufhin entsteht die finale Tabelle für alle Wortarten:

wort	lemma	wortart	animate	aspect	case	definitsness	degree	gender	negative	number	person	tense	type
znakove	N:znak;	N	N:n;		N:a;			N:m;		N:p;			
vreme	N:vreme;	N	N:n;		N:nav;			N:n;		N:s;			
vremena	N:vreme;N:vrijeme;	N	N:n;		N:gnav;			N:n;		N:sp;			
vremenu	N:vreme;N:vrijeme;	N	N:n;		N:dl;			N:n;		N:s;			
vremenom	N:vreme;N:vrijeme;	N	N:n;		N:i;			N:n;		N:s;			
vremenima	N:vreme;N:vrijeme;	N	N:n;		N:dil;			N:n;		N:p;			
vrijeme	N:vrijeme;	N	N:n;		N:nav;			N:n;		N:s;			
posao	N:posao;	N	N:n;		N:na;			N:m;		N:s;			
posla	N:posao;	N	N:n;		N:g;			N:m;		N:s;			

Abbildung 11: Ansicht der finalen Darstellung in MySQL

(10) Nun gilt es, zwei Masken anzulegen – eine für die Suche nach Informationen zur grammatikalischen Annotation und eine zweite für die Darstellung der Suchergebnisse.

(11) Der MorphoGenerator ist, wie bereits erwähnt, mit dem Gralis-Korpus verbunden und bietet informative Kanäle in zwei Richtungen – vom MorphoGenerator zum Korpus und vom Korpus zum MorphoGenerator, wobei das Ziel darin liegt, jedes Wort aus dem MorphoGenerator auch im Korpus darzustellen, um im Korpus Informationen über seine Häufigkeit und Umgebung sowie im Generator das Paradigma zu erhalten.

10. Wortarten

In der serbischen, kroatischen und bosni(aki)schen Sprache bildet ein Drittel der Wörter Substantive (37,43%), gefolgt von Adjektiven (32,34%) und Verben (29,89%). Diese drei Wortarten umfassen 92,48% des gesamten Wortschatzes (siehe Tabelle 3).

Sie besitzen ein reiches Flexionsparadigma, unterscheiden sich zum Teil hinsichtlich der Kategorien Genus, Kasus, Numerus, Tempus und Aspekt, sind belebt oder unbelebt, zeichnen sich durch ein umfassendes

Nr.	Wortart	Zahl der Wörter	%	Zahl der Regeln	Beziehung Zahl der Wörter – Zahl der Regeln
1	Substantive	37.606	37,43	311	0,0083
2	Adjektive	32.492	32,34	71	0,0022
3	Verben	30.030	29,89	378	0,0126
4	Numeralia	198	0,20	12	0,0606
5	Pronomina	135	0,13	50	0,3704
Insgesamt		**100.461**	**100,00**	**822**	**0,0052**

Tabelle 3: Die Häufigkeit der Wortarten und die morphologischen Annotationsregeln

Endungssystem aus und verfügen über verschiedene morphonologische Alternationen – vokalische (Jat-Reflex, Umlaut, Ablaut, Vokalisierung, Vokalausfall) und konsonantische (Assimilierung nach Stimmtonbeteiligung und Artikulationsort, Palatalisierung, Jotierung, unbeständiges a und e, fakultatives a, Konsonantenelision) – und zeigen prosodische Besonderheiten (Änderung des Akzentstelle, der Qualität und Quantität, postakzentuierte Längen u. Ä.).

Die durchgeführte Analyse im Untersuchungszeitraum verweist darauf, dass bei der Generierung von Wörtern die größte Zahl an Regeln für Verben (384) und Substantive (311) erforderlich ist, mit deutlichem Abstand gefolgt von Adjektiven (71), Pronomina (50) und Zahlwörtern (12). Bringt man die Regeln jedoch mit der absoluten Zahl an Wörtern in Verbindung, erhält man ein gänzlich anderes Bild: Für die morphologische Generierung der 135 pronominalen Wörter benötigt man mindestens 50 Regeln, für die 198 Numeralia 112 Regeln, während für die 37.606 Substantive 311 und für die 32.492 Adjektive bloß 71 Regeln ausreichen. Somit kann festgehalten werden, dass quantitativ eher abgeschlossene Wortarten (die in der Regel nicht um neue Wörter erweitert werden, wie hier die Pronomina und Numeralia) weitaus mehr Regeln notwendig machen als für Entlehnungen und Neubildungen offene Gruppen (Substantive und Adjektive). Die vorläufige Analyse zeigt, dass für die Generierung in prozentueller Hinsicht (in Bezug auf die Zahl der Regeln und die Zahl der Lexeme) die wenigsten Regeln für Adjektive (0,22%) und Substantive (0,83%) benötigt werden, weitaus mehr jedoch für Pronomina (37,04%)

und die meisten für Zahlwörter (56,57%) aufgestellt werden müssen. Der durchschnittliche Koeffizient für die Generierung aller Nomina liegt bei 0,77% (544 Regeln für 70.431 Substantive, Adjektive, Pronomina und Numeralia).

Die Regeln für die morphologische Generierung von Substantiven sind überaus komplex, da die Menge und die Heterogenität der Endungen, die Vielfalt der morphologischen Formen und die Komplexität der grammatikalischen Kategorien in Betracht zu ziehen sind. In den untersuchten Sprachen gibt es beinahe keine Substantive, die auf Grund einer Nullendung nicht zu generieren wären (dies betrifft nur *deci* ‚zehn Dekagramm', den weiblichen Namen *Miki* und eventuell noch einige andere Wörter). Die Zahl der unikalen Endungen ist nicht allzu groß. Sie beträgt neun (-a, -o, -e, -i, -u, -ø, -oj, -ama, -ima), doch wiederholen sich diese Endungen oft in unterschiedlichen Kasus: -i in elf, -a in acht, -e in sieben, -o, -u, -ama, -ima in drei, ø, -oj in zwei.

Bei der Generierung der Formen und ganzer Paradigmen stößt man auf unterschiedliche strukturelle Typen. Die bisherige Analyse lässt erkennen, dass zwei Extremfälle vorliegen, nämlich einer betreffend die Regeln für eine minimale Zahl (eins, zwei oder drei, z. B. deckt die Regel 80}^{vreoce} die zwölf sächlichen Substantive auf -e ab), wogegen im anderen Fall einige tausend nach gleichen Regeln strukturiert werden können (wie die Regel 03}^{strana}, mit der das Paradigma für 10.532 weibliche Substantive auf -a erzeugt wurde). Extremfälle bilden auch die Regeln für Substantive ohne vokalische oder konsonantische Alternationen, wie auch mit mehr oder weniger morphologischer Varianz. Es gibt auch Beispiele, in denen ein Wort eine gesonderte Regel verlangt, z. B. das Lehnwort *nargile* (Regel 280}^{nargile}). Es gibt auch Fälle, in denen ein Wort eine eigene Regel verlangt und dadurch einen eigenen morphologischen Typ bildet. In den Regeln kommt es zu Überschneidungen der siebengliedrigen Kategorie des Kasus (Nominativ, Genitiv, Dativ, Akkusativ, Vokativ, Instrumental und Lokativ), der viergliedrigen des Genus (maskulin, feminin, neutral und allgemein), der zweigliedrigen Kategorie der Zahl (Singular und Plural) und der Gegenüberstellung von belebt und unbelebt, wobei etwa unterschiedliche morphologische Typen maskuline Substantive für (1) die Belebtheitskategorie, (2) die Unbelebtheitskategorie und für (3) Wörter bilden, die beides zugleich ausdrücken.

11. Zusammenfassung

An der Karl-Franzens-Universität Graz wurde das Gralis-Korpus entwickelt, das für Analysen und das Erlernen aller slawischen Sprachen dient. Das Korpus besteht aus zwei Subsystemen – einem auditiven und einem textuellen. Das Text-Korpus enthält parallele Texte für Analysen zu allen slawischen Sprachen, wobei der Fokus auf der Befüllung mit Texten in südslawischen Sprachen und in der slawischen Sprache mit den meisten SprecherInnen (Russisch) lag. Im Gralis Text-Korpus gibt es drei Arten der Annotation: eine metatextuelle, eine extralinguistische und eine linguistische. Für die morphologische Annotation und die Analyse der grammatikalischen Struktur der Korpus-Sprachen dient das Online-Programm Gralis-MorphoGenerator. Für die morphologische Annotation wurde die Multext-East-Kodierung gewählt. Die Kodierung für das Gralis-Korpus besteht aus 20 Positionen. Die morphologische Annotation wurde bislang für 100.461 Wörter durchgeführt, wobei deren Paradigmen mit 822 Regeln generiert wurden.