UDK 811.112.2’366'373.74
811.163.42’366’373.74
Übersichtsartikel
Eingesandt am 16.10. 2006.
Angenommen für Publikation am 06.12. 2006.

Melita Aleksa
Josip-Juraj-Strossmayer Universität
Philosophische Fakultät
Abteilung für Germanistik
Osijek

Der kroatische HUMOR[1]: Überlegungen zu einer computergestützten morphologischen Analyse der flektierenden Sprachen[2]

Die vorliegende Arbeit beschäftigt sich mit der Problematik der computergestützten morphologischen Analyse der kroatischen Sprache im Vergleich zu einigen Aspekten der deutschen Problematik, aber auch mit den Fragen, die zum linguistischen Problembereich bei der Implementierung von HUMOR gehören. Es werden nicht die theoretischen Verfahren und die Arbeitsweise des Computerprogramms beschrieben, sondern die praktischen Lösungen, sprachwissenschaftlichen Anlässe und Überlegungen zum Adaptieren eines existierenden Parsers den flektierenden Sprachen. Die beschriebenen linguistischen Dilemmas umfassen die Bereiche des kroatischen Lexikons, der Sprachpolitik und der Fachliteratur.

Schlüsselwörter: HUMOR; Parser; Sprachpolitik; Sprachgeschichte; Flexionsparadigmen; lexikalische Basis; kroatisches Korpus.

1. Einführung

Die kroatische und die deutsche Sprache, den flektierenden Sprachen angehörend, sind zwei von mehreren Sprachen, die im Forschungsprojekt der morphologischen Analyse der Sprachsysteme mit Hilfe des Computer-

[1] High-speed Unification Morphology

[2] Die Arbeit wurde von der Magyar Ösztöndíj Bizottság - Stiftung unterstützt.

programms HUMOR beschrieben und untersucht werden bzw. wurden. HUMOR, oder High-Speed Unification Morphology ist ein von MorphoLogic entwickelter unifikationsbasierter morphologischer Parser, der in erster Linie der morphologischen Analyse von Sprachen dient. Außerdem stellt das Programm die Grundlage vieler computergestützter Übersetzungsprogramme dar. HUMOR wird unter anderem als Basis für sämtliche Übersetzungssysteme und Orthographieprüfer gebraucht, wie MobiMouse (ein Instrument zum Übersetzen von Wörtern, die auf dem Bildschirm angezeigt werden), MobiDic und MobiCat (Computerwörterbücher, die wechselseitig benutzt werden können) und MetaMorpho (ein Übersetzungssystem für das Übersetzten von einfachen Sätzen aus dem Englischen ins Ungarische und umgekehrt). Wie Prószéky und Kis in ihren wissenschaftlichen Publikationen schon bewiesen haben, besitzt das Programm vielerlei Anwendungsmöglichkeiten für die morphologische Untersuchung einer großen Anzahl verschiedener Sprachsysteme. Seine Vorteile kommen unter anderem durch die rasche Durchführung von Aufgaben zum Vorschein. Mit Hilfe von HUMOR sind bereits sowohl flektierende, als auch agglutinierende Sprachen untersucht worden, nämlich die ungarische, englische, deutsche und polnische Sprache.

Der vorliegende Aufsatz hat aber nicht die theoretische Beschreibung des Systems zum Ziel, sondern er beschäftigt sich in erster Linie mit der Problematik der Implementierung jenes morphologischen Analysators in ein flektierendes Sprachsystem, nämlich das Kroatische, im Vergleich zu einigen Aspekten der Problematik von deutscher Implementierung.

Im Gegensatz zu anderen Sprachen sind bei der Implementierung von HUMOR in das System der kroatischen Sprache neben linguistischen auch Schwierigkeiten anderer Art aufgetaucht, wie die Problematik der Sprachgeschichte und der Sprachpolitik. Das Ziel dieser Arbeit ist die Präsentation sowohl der auszuarbeitenden Version des Programms, als auch der aufgetauchten linguistischen Fragen und Dilemmas, die bis zu jener Phase des Forschungsprojektes in Erscheinung getreten sind, nämlich der Beschreibung verbaler, nominaler und adjektivischer Paradigmen des Kroatischen. Das Hauptziel des Forschungsprojektes ist nicht nur die Entwicklung der schon erwähnten Sprachtools für beide Sprachen, sondern auch das Konkretisieren der kroatischen Deklinations- und Konjugationsparadigmen, das sich als Hilfe für ein erfolgreicheres Lehren und Lernen des Kroatischen als Fremdsprache erweisen soll.

2. HUMOR

Die erste Demoversion des morphologischen Parsers HUMOR wurde von MorphoLogic im Jahre 1992 entworfen. Das Hauptziel lag nicht in der Entwicklung industrieller Orthographieprüfer, Wortrennungsprogramme und Thesauri, da solche Programme schon seit Jahren auf dem Arbeitsmarkt vertreten sind, sondern in erster Linie im linguistischen Parsen von Sprachen für verschiedene Suchzwecke und das flache bzw. volle Parsen in übersetzungsunterstützenden Systemen (Prószéky und Kis 1999: 266). Nach Prószékys Worten lag die erste Absicht beim Entwerfen eines morphologischen Analysators im Bedürfnis nach der Sammlung möglichst vieler Informationen über ein bestimmtes Wort. Die Zweite lag in der Implementierung des Parsers selbst (Prószéky und Kis 1999: 267). Das bedeutet, dass Informationen auf verschiedenen linguistischen Ebenen gesammelt werden, woraus später ein HUMORsches Lexikon kompiliert wird. In erster Linie dient HUMOR dem linguistischen Stemming, Verbesserung der Orthographie und den Vorarbeiten beim Parsen der Lemmata (Prószéky und Kis 2002: 3). HUMOR wird, wie schon erwähnt, als Basis für andere Übersetzungsprogramme benutzt, wie MobiDic Computerwörterbücher, die heutzutage erfolgreich für die englische, deutsche und ungarische Sprache angewandt werden. Die experimentellen Versionen von MobiDic schließen auch Spanisch, Polnisch und Japanisch ein (Prószéky und Kis 2002: 3). Weiter dient HUMOR auch als Basis für MetaMorpho, ein System für das Übersetzen einfacher Sätze aus dem Englischen ins Ungarische, und MobiMouse, ein Programm, das Übersetzungsmöglichkeiten von Wörtern bietet, die auf dem Bildschirm angezeigt werden.[3]

Außer HUMOR gibt es heutzutage mehrere andere morphologische Parser. Vielleicht haben sich als die bedeutendsten die Programme im Rahmen von XEROX erwiesen (XFST, TWOLC, LEXC) (Beesley–Kartunnen 2003). Außer diesen gibt es auch andere ähnliche Programme, geeignet u.a. für das morphologische Parsen der deutschen Sprache, wie GERTWOL[4] und MORPHY. [5]

---

[3] Laut Prószéky und Kiss, "[t]he tools that use HUMOR are described as context–sensitive instant comprehension tools, more than a dictionary lookup engine as they tailor dictionary entries to the context of the translation point. The tool is less than a translation engine, as it performs no syntactic processing of the source text, only a series of dictionary lookups". (Prószéky und Kis 2002).

[4] http://www2.lingsoft.fi/cgi-bin/gertwol.

[5] http://wordnet.princeton.edu/man/morphy.7WN.html#toc0.

Im Rahmen meiner Doktorarbeit und zugleich des Forschungsprojektes, arbeite ich schon seit drei Jahren zusammen mit MorphoLogic an der Entwicklung vom kroatischen HUMOR. Das Projekt, wie schon erwähnt, hat die Implementierung des erwähnten Parsers in das System der kroatischen Sprache zum Ziel, was weiterhin zu anderen Anwendungsbereichen führt. Als Ergebnis der bisherigen Arbeiten kann die Tatsache hervorgehoben werden, dass der Parser ungefähr 80% der Wörter aus einem Text erkennt und fähig ist, sie morphologisch zu analysieren. Das endgültige Ziel des Projektes impliziert sowohl die Möglichkeit der Entwicklung der oben genannten Programme im Sinne von kroatisch-deutschen, aber auch anderen Verbindungen, als auch die Entstehung und Veröffentlichung einer morphologischen Datenbank, die eine zusätzliche Hilfe für das Lernen des Kroatischen anbieten wird.

2.1. *Die Funktionsweise des Programms*

Um die Funktionsweise des Programms zu verdeutlichen, sind einige Konzepte in Bezug auf die Analyse selbst am Beispiel des Kroatischen zu klären. Weil die Prozesse des morphologischen Parsens unter der Oberfläche verlaufen, unterscheidet sich der Output im Wesentlichen von den Verlaufsergebnissen auf der Subebene, was auch im folgenden Beispiel zu sehen ist. Dem Programm wurde der Befehl gegeben, die Form *damo* ‘wir geben’ (*1. Person Plural)* des kroatischen Verbs *dati* ‘geben’ zu analysieren, was zu folgendem Ergebnis führte:

Analysis of “damo”:

(1) dama[Sf]=dam+o[51]

(2) dati[Vs]=da+mo[p1]

Das Programm hat die gegebene Form mit zwei verschiedenen Lemmata verbunden, dem Verb *dati* und dem Substantiv *dama* 'die Dame'. In (1) kann man sehen, dass *damo* eine Vokativ Singular Form [51] des angeführten Substantivs ist, während in (2) die gegebene Form mit der 1. Person Plural Präsens [p1] kombiniert wurde.

Dieses Beispiel kann auch als ein typisches Stemmingproblem bezeichnet werden, da ein Suchbefehl mit verschiedenen Lemmata verbunden wird. Laut Prószéky und Kis (1999) liegt die Entscheidung beim Endbenutzer selbst, oder an einem kontextbasierenden Disambiguator (Prószéky und Kis 1999: 267).

2.2. *Das kroatische Lexikon*

Außer der Engine, das als Triebwerk des Programms bezeichnet werden kann, spielt beim Ablauf des Programms auch die zweiteilige lexikalische Basis eine wichtige und entscheidende Rolle. Die im Lexikon kompilierte Basis der kroatischen Sprache enthält 60 000 lexikalische Einheiten, die aus Anić's (2000) *Rječnik hrvatskoga jezika (Wörterbuch der kroatischen Sprache)*[6] stammen. Es ist vielleicht wichtig zu betonen, dass die Anzahl der lexikalischen Einheiten nicht begrenzt ist, d.h. neue Lemmata können jederzeit dem Lexikon beigefügt bzw. dupliziert werden. Der zweite Teil des Lexikons besteht aus der Beschreibung von Deklinations- und Konjugationssystemen, die aufgrund folgender Werke angefertigt wurde: Barić et.al. (1995) *Hrvatska gramatika (Kroatische Grammatik),* Raguž (1997) *Praktična hrvatska gramatika (Praktische kroatische Grammatik),* Silić und Pranjković (2005) *Gramatika hrvatskoga jezika za gimnazije i visoka učilišta (Kroatische Grammatik für Gymnasien und Hochschulen)*, Težak und Babić (2005) *Gramatika hrvatskoga jezika: priručnik za osnovno jezično obrazovanje (Kroatische Grammatik: ein Handbuch für die elementare Sprachausbildung)* und Težak's Werke (1991, 1995, 1999, 2000). Die orthographischen Regeln wurden Babić, Finka und Moguš's (1996) *Hrvatski Pravopis (Kroatische Rechtschreibung)* entnommen. Über Entscheidungen und Probleme, die diese Werke betreffen, wird später noch die Rede sein.

Das Hauptprinzip beim morphologischen Parser HUMOR, das auch als Grundlage für den grammatischen Input bezeichnet werden kann, liegt in der Aufteilung der Lemmata auf *Stems* und *Terms*. Die traditionellen morphologischen Kategorien wie Wortstamm und Affixe werden hier absichtlich nicht benutzt, da die im Parser definierten Begriffe mit denen nicht immer übereinstimmen.[7] Allgemein gesagt umfasst die Kategorie von S*tems* jene Wortteile, die im Laufe der Flexion unverändert bleiben, während der veränderliche Rest des Lemmas als *Term* bezeichnet werden kann. Wichtig zu betonen ist auch die Tatsache, dass ähnlich wie bei der traditionellen Morphologie ein Wort aus keinem Ø–*Stem*, aber aus einem Ø–*Term* bestehen

[6] Im folgenden: Wörterbuch.

[7] Laut Prószéky und Kis, "[c]oncatenation of stem allomorphs and suffix allomorphs is licensed with the help of the following two factors: continuation classes defined by paradigm descriptions and classes of surface allomorphs. The latter is a cross–classification of the paradigms according to phonological and graphemic properties of the surface forms. Both verbal and nominal stem allomorphs can be characterized by sets of suffix allomorphs that can follow them. When describing the behaviour of stems, all suffix combinations beginning with the same morpheme are considered equivalent because the only relevant pieces of information come from the suffix that immediately follows the stem" (Prószéky & Kis 1999: 262).

kann. Am Beispiel des kroatischen Wortes *brzina* 'die Geschwindigkeit' ist zu sehen, dass im Gegensatz zu den traditionellen Kategorien *brzin-* als ein Stem bezeichnet wird, gefolgt von einem Term *–a*, weil der Stem-Wortteil während der Flexionen unverändert bleibt. Die auf der vorläufigen Kategorisierung basierenden Prozesse bleiben unter der Oberfläche, während das Output andere Informationen beinhaltet, was am folgenden Beispiel zu sehen ist:

Analysis of "stranke"

(3) stranka[Sf]=stran+ke[21]

(4) stranka[Sf]=stran+ke[12;42;52]

Wie aus diesem Beispiel zu sehen ist, hat das Programm bei der morphologischen Analyse die Form *stranke* 'Parteien' mit mehreren Erscheinungsformen des Lemmas *stranka* in Verbindung gesetzt. Diese sind Genitiv Singular [21], Nominativ Plural [12], Akkusativ Plural [42] und Vokativ Plural [51]. Die Terms sind vom Stem (*stran-)* mit einem + Zeichen abgegrenzt. Die erste Zahl, wie schon erläutert, bezeichnet den Kasus, die Zweite das Genus.

2.3. *Das kroatische Korpus*

Damit die Daten verifiziert werden können, wurde das ganze System an einem aus 50 Millionen Wörtern bestehenden Korpus geprüft. HUMOR ist nämlich ein partiell sich selbst korrigierendes System. Das bedeutet, dass Fehler anhand der Korpusanalyse entdeckt werden, aber diese müssen dann manuell verbessert werden. Um die Verifikation der Daten zu unterstützen, braucht man ein repräsentatives Korpus, das die meisten Bereiche der Sprache umfasst und die Sprachsituation möglichst genauer schildert.

Das für dieses Projekt zusammengestellte kroatische Korpus besteht aus einer gewissen Anzahl von Subkorpora. Die verwendeten Werke stammen aus den Bereichen der Literatur, den Schriftmedien, dem Internet, und den Fach- und Lehrbüchern. Bei der Zusammenstellung des Testkorpus sind auch einige spezifische Probleme aufgetaucht, die teilweise auch die Sprachpolitik betreffen.

Zu dem Bereich der literarischen Werke, die im Testkorpus vorhanden sind, wurden repräsentative Werke aus den drei Gattungen Lyrik, Drama und Epik gewählt. Die Auswahl bezüglich ihrer Repräsentativität wurde der schon existierenden Kompilation der wichtigsten kroatischen literarischen Werke überlassen, nämlich der Bulaja (1999, 2000, 2002) *Klasici hrvatske književnosti*

(Klassiker der kroatischen Literatur), die ein kroatisches Opus vom 16. bis zum 20. Jahrhundert enthält. Weil ein Testkorpus möglichst viele Bereiche der Sprache umfassen sollte (darunter neben der Schrift- auch die gesprochene Sprache), wurden auch der gesprochenen Sprache nahe stehende Gattungen gewählt, nämlich das Drama und die Blogs. Da die erwähnte Kompilation literarischer Werke die neuesten Ereignisse nicht enthält, wurde der Mangel an Kodifizierungen der gesprochenen Sprache durch den Bereich des Dramas und die von Tippfehlern gereinigten Texte von Internet-Blogs ersetzt. Zu einem weiteren Subkorpus gehören Texte, die aus den aktuellen Schriftmedien entnommen sind, und zwar aus mehreren Bereichen: aus verschiedenen Zeitschriften (Fachzeitschriften, Jugend- und Kindermagazine) und Zeitungen, während andere Subkorpora, wie schon erwähnt, aus Texten der Fach- und Lehrbücher bestehen.

3. Linguistische Fragen

3.1. *Auswahl der Sprachvarianten*

Im Vergleich zur linguistischen Problematik bei der Implementierung von HUMOR ins Sprachsystem der deutschen Sprache umfassen die linguistischen Fragen, die sich auf das Kroatische beziehen, die Auswahl der Sprachvariante, die Sprachpolitik und die traditionellen Beschreibungen der Konjugations- bzw. Deklinationssysteme. Der größte Anteil der linguistischen Dilemmas im Deutschen verweist auf die Problematik der Rechtschreibung.

Nachdem die Buchstaben der entsprechenden Sprache bestimmt wurden, mussten einige wichtige Fragen (Fragen der Sprachvariante und der Sprachkode) beantwortet werden. Das erste Problem war die Inkorporierung der beiden Sprachvarianten, d.h. sowohl der Schrift- als auch der gesprochenen Sprache. Aufgrund des mangelnden Korpus der kodifizierten gesprochenen Sprache wurde jedoch darauf verzichtet. Ein weiterer Grund dafür ist, dass die genaue phonetische Kodifizierung aller Varianten der gesprochenen Sprache im Kroatischen nicht existiert, d.h. bis heute nicht kompiliert wurde. Weitere Motive, die gegen die Beachtung der gesprochenen Sprache genannt werden können, sind die mangelhafte Kodifizierung der dialektalen und regionalen Sprachvarianten sowohl in der Schrift-, als auch in der gesprochenen Sprache. Damit wurde zugleich ein weiteres Problem gelöst, und zwar die Auswahl zwischen der Standardvariante oder dem Miteinbeziehen der regionalen Sprachvarianten. Aus allen diesen Gründen stellt man fest, dass die Beschreibung der kroatischen Sprache in HUMOR sich auf die geschriebene Standardvariante bezieht.

Eine weitere Problematik in Bezug auf die gewählte Variante des Kroatischen verweist auf die Fragen der Sprachpolitik.

3.2. *Die kroatische Sprachpolitik*

Die kroatische Sprache, den anderen südslawischen Sprachen ähnlich, hatte einen langen Kodifizierungsprozess, der schon im 16. Jahrhundert begonnen hat. Versuche, ein unifiziertes graphologisches System zu gestalten, gab es bis zum 19. Jahrhundert. In diesem Zeitraum ist eine große Anzahl an verschiedenen Varianten der Phoneme zu sehen, was ein ernstes Problem für die morphologische Analyse der kroatischen Sprache darstellt. Die Mehrheit der bis zum 19. Jahrhundert benutzten Grapheme kann man auf folgende Weise zusammenfassen (Moguš 1995):

Grapheme der heutigen kroatischen Sprache	Einige Grapheme, die bis zum 19. Jh. benutzt wurden
č	*ç, ʒ́, cs, ts, cz*
ć	*c', ch, tj*
đ	*gh, dy*
lj	*l̊, ļ, ly, gl*
nj	*ñ, ń, nj, ny, gn*
š	*ſ, sh,*
ž	*ʒ, sh, x*
-je /- ije	*ě*

Abb.1. Grapheme des Kroatischen heute und im 19. Jahrhundert (Moguš 1995)

Da die einzige existierende Variante des slawischen HUMOR, die Polnische, auch die Verarbeitung von Texten aus dem 18. Jahrhundert ermöglicht, entstand ein solcher Bedarf auch bei der kroatischen Sprache. Wenn man die oben angeführten Grapheme und die Sprachsituation in der Zeit betrachtet, kann man feststellen, dass eine solche Entscheidung zu meist komplizierten, teils auch unmöglichen Ergebnissen in Bezug auf die Anzahl und Entstehung der Stems und Terms führen würde. Deswegen wurde die Entscheidung getroffen, nur die Sprache zu beschreiben, die nach dem abgeschlossenen Prozess der Kodifizierung entstand (ab 1830), was aber wiederum zu anderen Fragen führte.

Die kroatische Sprache, diachronisch gesehen, kann grob in drei Perioden aufgeteilt werden, die Zeit bis zum 19. Jahrhundert mit der nichtunifizierten Graphologie, dann die Periode des Serbokroatischen, und schließlich die Zeit ab 1990, die Zeit der heutigen, modernen kroatischen Sprache. Unter diesen drei Sprachvarianten bestehen Unterschiede auf fast allen Sprachebenen, was zu einer Problematik der Zusammenstellung der kroatischen lexikalischen Basis führt. Wenn beschlossen worden wäre, eine morphologische Analyse älterer Sprachvarianten zu ermöglichen (z.B. des Serbokroatischen), müsste man auch die Lexeme aus beispielsweise der serbischen Sprache in das Lexikon inkorporieren. Die Frage ist, ob dieses Lexikon dann noch als kroatisches Lexikon gelten würde. Im anderen Fall, wenn man für eine lexikalische Basis nur die Lemmata aus Anić's *Rječnik hrvatskoga jezika* nehmen würde, würden sich dann Schwierigkeiten bei der Korpusanalyse ergeben, und das Programm würde unfähig sein, die vor den 1990er Jahren entstandenen, in serbokroatischer Sprache geschriebenen Werke zu analysieren. Das Hauptproblem liegt darin, dass auch heute serbische Lexeme einen Bestandteil des kroatischen gesprochenen Lexikons ausmachen, und dass diese in großem Maße auch auf kroatischen Internetseiten vertreten sind. In einer Internetuntersuchung wurde das serbische Wort *ponekad* (Brodnjak 1991:411) mit dem kroatischen Wort *katkad* 'manchmal' verglichen. Es hat sich herausgestellt, dass das serbische Lexem auf ca. 110 000 ausschließlich kroatischen Internetseiten vorkommt (Google 2005), was auf Spuren aus der serbokroatischen Sprache hinweist.

Im Gegensatz zur kroatischen Sprache liegen die meisten Schwierigkeiten bei der automatischen morphologischen Untersuchung der deutschen Sprache in der Rechtschreibung. Die deutschen Rechtschreibreformen haben einige Konsequenzen in Bezug auf HUMOR mit sich gebracht, was in der Vermehrung der möglichen Stems und Terms zu sehen ist. Ähnlich wie beim Kroatischen, wenn man ermöglichen will, dass das Programm auch ältere oder neuere Texte analysieren kann, muss man die neuen Varianten der Wörter dem Lexikon hinfügen. Ein weiteres technisches Problem liegt in der Tatsache, dass das Leerzeichen in HUMOR die Analysebegrenzungen bedeutet. Diesbezüglich muss man andere Lösungen finden, um diese Grenzen zu überwinden.

3.3. *Fachliteratur und grammatische Regeln*

Bis zum jetzigen Stand der Forschung, nämlich der Beschreibung verbaler, nominaler und adjektivischer Paradigmen der kroatischen Sprache mit Hilfe von HUMOR, sind einige Mängel der erwähnten Fachliteratur zum Vorschein gekommen.

Die erste bedeutende Frage ist die Anzahl der Wörter, die in Anić's (2000) Wörterbuch angeführt sind. Obwohl das Wörterbuch 60 000 Wörter der kroatischen Sprache zählt, sind es weitaus weniger, da einige dupliziert oder tatsächlich nicht zu den gebrauchten Lemmata gezählt werden können. Diese sind z.B. die nominalisierten Buchstaben, wie *S, n.* Die tatsächliche Anzahl der Wörter liegt bei 56 000.

Das zweite Problem zeigt sich in der Auseinandersetzung mit den orthographischen Regeln. Beim Adjektiv *ungarisch* beispielsweise sind laut Babić, Finka und Moguš (1996) zwei orthographische Formen erlaubt, nämlich *mađarski* und *madžarski*. Da im Wörterbuch nur eine Form vertreten ist, musste die andere in die lexikalische Basis von HUMOR manuell nachgetragen werden.

Die Hauptprobleme bei der Beschreibung grammatischer Paradigmen beziehen sich zum einen auf die ungenauen Beschreibungen der Flexionen, und zum anderen auf die mangelhaften Darstellungen der Flexionsregeln. Bei Verben sind diese in den nicht konkretisierten Beschreibungen ihrer Formen zu sehen, besonders wenn man *aorist* und *imperfekt* betrachtet (Aleksa 2006a: 262). Außerdem gibt es einige Auseinandersetzungen mit dem Wörterbuch, besonders wenn vom Gerundium und von adjektivisierten Verbformen die Rede ist. (vgl. Aleksa 2006b). Bei den Beschreibungen nominaler Paradigmen gibt es Unklarheiten die Genera betreffend, während die Beschreibungen adjektivischer Paradigmen in der entsprechenden Fachliteratur auch auf gewisse Fragen hindeuten (vgl. Aleksa 2006b: 11).

4. Fazit und Ausblick

Bei der Implementierung des morphologischen Parsers HUMOR ins kroatische Sprachsystem sind einige Fragen aufgetaucht, die sowohl die Sprachwissenschaft, als auch die Sprachpolitik betreffen. Alle beschriebenen Dilemmas sind bis zum derzeitigen Stand des ganzen Projektes zu entdecken, nämlich der Beschreibung nominaler, verbaler und adjektivischer Paradigmen. Das Hauptziel des ganzen Projektes ist, wie schon erwähnt, außer der Entwicklung einer kroatischen Version von HUMOR auch ihre Verbindung mit der schon existierenden deutschen Version, wobei als endgültiges Ziel die Entwicklung anderer computergestützter Übersetzungsprogramme gesehen werden kann. Die Fortsetzung des ganzen Projektes ist die Beschreibung weiterer Flexionsparadigmen. Die neuen Problembereiche werden aber das Thema anderer Arbeiten sein.

Literatur

Aleksa, Melita (2006a). A HUMOR rendszer adaptálása a horvát nyelvre. Klaudy, Kinga, Csilla Dobos, Hrsgg. *A világ nyelvei és a nyelvek világa. MANYE XV*.(2/1). Pécs—Miskolc, MANYE, 259-264.

Aleksa, Melita (2006b). Automatic morphological analysis of the Croatian language - The verbal, adjectival and nominal inflections within the morphological parser HUMOR. *CESCL1- Proceedings of the First Central European Student Conference in Linguistics*. http://www.nytud.hu/cescl/proc.html.

Anić, Vladimir (2000). *Rječnik hrvatskoga jezika*. Zagreb: Novi Liber.

Babić Stjepan, Božidar Finka, Milan Moguš (1996). *Hrvatski pravopis.* Zagreb: Školska knjiga.

Barić, Eugenija. et. al. (1995). *Hrvatska Gramatika*. Zagreb: Školska Knjiga.

Beesley, Kenneth R., Lauri Karttunen (2003). *Finite State Morphology*. http://www.stanford.edu/~laurik/fsmbook/home.html, gesehen am 06.12.2006.

Brodnjak, Vladimir (1991). *Razlikovni rječnik srpskog i hrvatskog jezika*, Zagreb: Školske novine.

Elekfi László (1994). *Magyar ragozási szótár*. Budapest: MTA Nyelvtudományi Intézete.

Moguš, Milan (1995). *Povijest hrvatskoga književnoga jezika*. Zagreb: Globus.

Moguš Milan, Maja Bratanić, Marko Tadić (1999). *Hrvatski čestotni rječnik*. Zagreb: Školska knjiga – Zavod za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu.

Papp Ferenc (1969). *A magyar nyelv szóvégmutató szótára.* Budapest: Akadémiai Kiadó.

Prószeky, Gábor, Balász Kis (1999). A unification-based approach to morpho-syntactic parsing of agglutinative and other (highly) inflectional languages. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics.* Maryland, USA: College Park, 261-268. http://www.morphologic.hu/h_pgpub.htm, gesehen am 2. August 2006.

Prószéky, Gábor, Balázs Kis (2002). Context-Sensitive Dictionaries. *COLING-2002,* Taipei, Taiwan. San Francisco: Morgan Kaufman, http://portal.acm.org/citation.cfm?id=1071899, gesehen am 2. August 2006.

Raguž, Dragutin (1997). *Praktična hrvatska gramatika.* Zagreb: Medicinska naklada.

Silić, Josip, Ivo Pranjković (2005). *Gramatika hrvatskoga jezika za gimnazije i visoka učilišta*. Zagreb. Školska knjiga.

Tadić, Marko (1994). *Računalna obrada morfologije hrvatskoga književnog jezika*. Ph.D. Thesis, Manuscript. Zagreb: Filozofski fakultet Sveučilišta u Zagrebu. http://www.hnk.ffzg.hr/mt/, gesehen am 2. August 2006.

Težak, Stjepko (1991). *Hrvatski naš svagda(š)nji.* Zagreb: Školske novine.

Težak, Stjepko (1995). *Hrvatski naš osebujni*. Zagreb: Školske novine.

Težak, Stjepko (1999). *Hrvatski naš (ne)zaboravljeni*. Zagreb: Tipex.

Težak, Stjepko (2004). *Hrvatski naš (ne)podobni*, Zagreb: Školske novine.

Težak, Stjepko, Stjepan Babić (2005). *Gramatika hrvatskoga jezika: priručnik za osnovno jezično obrazovanje*. Zagreb: Školska knjiga.

Wardhaugh, Roland (1995). *Szociolingvisztika.* Budapest: Osiris-Századvég.

Adresse der Autorin:

Josip-Juraj-Strossmayer Universität
Philosophische Fakultät
Abteilung für Germanistik
L. Jägera 9
HR-31000 Osijek
maleksa@ffos.hr

HRVATSKI HUMOR:
PROBLEMATIKA RAČUNALNE MORFOLOŠKE ANALIZE FLEKSIJSKIH JEZIKA

U radu se progovara o problematici automatske morfološke analize hrvatskoga i njemačkoga jezika, kao i o lingvističkim pitanjima koja su se pojavila prilikom prilagodbe već postojećeg morfološkoga analizatora (nazvanog HUMOR) sustavu hrvatskoga i njemačkoga jezika. U radu se ne predstavlja teoretski način rada samoga programa, već se naglasak stavlja na lingvistička pitanja koja su se pojavila prilikom provođenja projekta automatske morfološke analize fleksijskih jezika. Opisani problemi obuhvaćaju pitanja sastavljanja leksičke baze programa, jezične politike i dileme u glede stručne literature.

Ključne riječi: HUMOR; parser; jezična politika; povijest jezika; deklinacijska i konjugacijska paradigma, leksička baza, hrvatski jezični korpus.