BESIM KABASHI

Zeichen für Gjon Buzuku. Die Zusammenarbeit zwischen der albanischen Linguistik und der Computerlinguistik*

1 Einleitung

Die Computerlinguistik entwickelt Modelle und Methoden, um die maschinelle Verarbeitung natürlicher Sprache zu ermöglichen. Beispiele dafür sind die automatische Erkennung gesprochener Sprache und die automatische Analyse eines Satzes.

Die Zusammenarbeit der Computerlinguistik mit der traditionellen Linguistik ist vielfältig: Die Computerlinguistik übernimmt aus der Linguistik bewährte Methoden und Erkenntnisse. Im Gegenzug unterstützt die Computerlinguistik die Linguistik, indem sie Techniken und Programme entwickelt, welche die Benutzung der linguistischen Daten vereinfachen und deren Weiterverarbeitung ermöglichen. Als Beispiel wären hier Konkordanzprogramme und Auszeichnungssprachen (*Markup Languages*) zu nennen. Überdies erlaubt es die Computerlinguistik, linguistische Theorien zu überprüfen und zu evaluieren.

Im Folgenden werden einige Themen aus der Albanistik und der Computerlinguistik, wie Textkodierung, insbesondere von Texten aus älteren Sprachstufen, Korpora und lexikographische Werke gestreift. Im Anschluss werden computerlinguistische Programme vorgestellt und einige Vorschläge zur Modellierung linguistischen Wissens unterbreitet.

2 Zeichensätze für ältere Texte: Zeichen für Buzuku

Alle Buchstaben eines Alphabets und weitere Zeichen, wie Ziffern und Interpunktion, bilden zusammen einen Zeichensatz, für dessen technische Umsetzung, das heißt Zuweisung eines numerischen Wertes für jedes Zeichen, verschiedene Standards existieren. Es gibt zahlreiche Zeichensätze, die auf die jeweiligen Alphabete verschiedener Sprachen zugeschnitten sind.

---

* Für den Vorschlag und die Motivation, dieses Thema in diesem Kreis zu behandeln, sei B. Demiraj (München) gedankt. Ich bedanke mich auch bei den Teilnehmern dieser Tagung und bei meinen Kollegen (Universität Erlangen-Nürnberg) für zahlreiche Fragen und Diskussionen zu diesem Thema.

Das heutige Albanisch wird von einigen Zeichensätzen abgedeckt, unter anderem von dem für die westeuropäischen Sprachen vorgesehenen *Latin-1*-Zeichensatz (*International Organization for Standardization, Norm 8859-1*, kurz ISO-8859-1). Die Verarbeitung altalbanischer Texte ist jedoch insofern ein Problem, als diese Grapheme enthalten, die nicht in Latin-1 vorkommen und daher durch andere, ähnliche Zeichen dargestellt werden müssen.

Eine größere Auswahl an Zeichen bietet der Zeichensatz *Unicode*. Wie ISO-8859-1 ordnet er jedem Zeichen eine eindeutige Nummer zu, beschränkt sich aber nicht auf die 256 Zeichen, die mit 8 Bit kodiert werden können sondern setzt sich die Kodierung aller benötigten Zeichen aus verschiedenen Sprachen, Sprachstufen und Schreibsystemen in einem einzigen System zum Ziel.[1] So werden verschiedene Kodierungssysteme und die daraus resultierenden Probleme vermieden.

Es gibt Zeichen, wie z. B. das ƕ von *Buzuku* (ÇABEJ 1987 und RESSULI 1958), die in Unicode (bis zur jetzigen Version 4.1.0) noch nicht aufgenommen wurden und daher nicht mittels Unicode kodiert werden können. Diese Zeichen müssen durch ähnliche Zeichen repräsentiert werden. Beispielsweise wird anstatt des Zeichens ƕ das ħ[2] verwendet. Ersetzungen (Transliterationen) dieser Art erscheinen wenig sinnvoll, vielmehr ist die Aufnahme dieses und anderer Zeichen (ƕ, ɮ und ɤ bei *Buzuku*) in Unicode wünschenswert.[3] Das Ziel sollte die Erstellung von Zeichen sein, die das Original so genau wie möglich graphisch repräsentieren. Altalbanische Texte könnten dann leichter reproduziert werden.

Durch Ersetzung von Zeichen oder Wechsel des Zeichensatzes innerhalb einer Datei kommt es bei der maschinellen Sprachverarbeitung, beispielsweise bei der Textreproduktion, oft zur Verfälschung des Ergebnisses. Dennoch kann bis zur Aufnahme in Unicode zur originalgetreuen

---

[1] UTF-8 ist eine der dem Unicode-Standard entsprechenden Kodierungen variabler Länge (1-4 Byte, also von 1x8 bis 4x8 Bit). Damit können alle Zeichen, die Unicode definiert, abgedeckt werden. Für ausführliche Informationen über ISO und Unicode siehe *http://www.iso.org* beziehungsweise *http://www.unicode.org* .

[2] *Latin small letter h with stroke*, Unicode-Zeichennummer U+0127 aus dem Bereich (*Range* 0100-017F) / aus der Tabelle *Unicode Latin Extended-A*.

[3] Hierzu wäre eine Initiative zu ergreifen, diese Zeichen zu identifizieren, zu sammeln, technisch aufzubereiten und einen Antrag auf Aufnahme in Unicode zu stellen. Eine Kodierung in Unicode sollte alle Zeichen enthalten, die in (alt)albanischen Texten enthalten sind. Auch die Digraphe *dh, gj, ll, nj, rr, sh, th, xh* und *zh* sollten berücksichtigt werden. Dies wäre in erster Linie die Aufgabe einer *Nationalen Organisation für Standardisierungen* oder der *Akademie der Wissenschaften der Republik Albanien* bzw. *Kosovas Akademie der Wissenschaften und Künste*.

Textreproduktion ein proprietärer Zeichensatz für Altabanisch verwendet werden.[4]

3 Linguistische Daten und linguistisches Wissen: Korpora und Lexika

Für die maschinelle Sprachverarbeitung werden linguistisches Wissen und sprachspezifische Daten, insbesondere Korpora, Lexika und Klassifikationssysteme, benötigt.

3.1 Korpora

Ein Korpus dient als Referenz beziehungsweise Informationsquelle, erleichtert linguistische Untersuchungen und stellt sie auf eine empirische Grundlage. Korpora können einerseits für die Überprüfung von Sprachtheorien und -hypothesen und andererseits zum Testen und Trainieren von sprachtechnologischen Systemen eingesetzt werden. Ein Korpus ist für fundierte linguistische Untersuchungen eine *conditio sine qua non*.

Für das Altalbanische gibt es einen vielversprechenden Ansatz: Das von F. Altimari an der Università della Calabria geführte *Archivio Letterario / Arkivi Letrar*. Ein gutes Beispiel ist auch das *Titus-Projektes* (*Thesaurus Indogermanischer Text- und Sprachmaterialien*) der Universität Frankfurt, in dessen Rahmen der Text des *Meshari*, bereitgestellt von W. Hock (Berlin), gefunden werden kann.[5]

Die Methoden und Programme der Korpuslinguistik können auf verschiedene Sprachen angewendet beziehungsweise mit wenig Aufwand an diese angepasst werden. Dennoch gibt es bis zu diesem Zeitpunkt nur ein kleines Korpus für das heutige Albanisch, das als einfacher Text in ECI/MCI (1994) enthalten ist. Dies sollte die 2004 in Osnabrück gestartete *International initiative for a reference corpus of Albanian*, kurz *IIRCA*, unter Leitung von J.-L. Duchet (Poitiers) und R. Përnaska (Tiranë / Paris) ändern.

Bei der Konzeption von Korpora ist vor allem darauf zu achten, dass diese unabhängig vom verwendeten Betriebssystem nutzbar, beispielsweise webbasiert, sind, dass verschiedene normierte Datenaustauschformate zur Verfügung gestellt werden und dass das Konzept für alte und moderne Texte gleichermaßen geeignet ist. Zum Teil wurden diese Desiderata in *IntraText*[6] umgesetzt.

---

[4] Der Zeichensatz wurde von B. Kabashi für Linux, MacOS X und Microsoft Windows erstellt. Siehe hierfür *http://www.linguistik.uni-erlangen.de/~bmkabash/buzuku.html* .

[5] *http://titus.uni-frankfurt.de/texte/etcs/alban/buzuku/buzuk.htm*

[6] Unter *http://www.intratext.com* sind einige albanische Texte zu finden, samt der Möglichkeit, Konkordanzen und verschiedene Statistiken zu erstellen.

3.2 Lexika

Lexika sind für die maschinelle Analyse und Produktion natürlicher Sprache unabdingbar und bilden den Kern der maschinellen Sprachverarbeitung. Benötigt werden vor allem Lexika mit umfangreichen Angaben verschiedenster Art, von der Rechtschreibung über Morphologie und Syntax bis hin zur Semantik.

Das albanische Rechtschreibwörterbuch (FJDSH 1976) ist mit fast 30 Jahren zu alt, insbesondere in Anbetracht der sozio-politischen Veränderungen der albanischen Gesellschaft in den letzten 15 Jahren. Diese haben einen großen Einfluss auf die Sprache ausgeübt, vor allem im Bereich der Lexik. Als zeitliche Überbrückung aus computerlinguistischer Sicht und als notwendige Ergänzung dienen die Werke *Rückläufiges Wörterbuch der albanischen Sprache* (SNOJ 1994) und das *Inverse Dictionary of Albanian*[7] von A. Murzaku. Das albanische Standardwörterbuch (FJALORI 1984) enthält viele grammatische Angaben, wurde aber in der zweiten Auflage (FJALORI 2002) nicht hinreichend aktualisiert. Die Entwicklungen der letzten Jahre in der albanischen Lexikographie geben dennoch Anlass zur Hoffnung: Erfreulich ist das Erscheinen eines Antonymwörterbuchs (SAMARA 1998), eines Phraseologischen Wörterbuches (THOMAI 1999) und zweier Synonymwörterbücher (DHRIMO ET AL. 2002 und FJS 2005), die unter anderem für eine computerlinguistische Modellierung von Semantik bedeutsam sind.

4 Wichtige Aufgaben für die Zukunft

Neben der Aufnahme der albanischen Zeichen in Unicode sowie der Erstellung von Korpora und Lexika haben aus der Sicht der Computerlinguistik vor allem die folgenden Punkte höchste Priorität:

(1) Eine detaillierte Systematisierung der morphologischen Phänomene. Insbesondere eine Klassifikation des albanischen Nominalsystems für jeden Lexikoneintrag ist wünschenswert. Das Konjugationssystem von BUCHHOLZ ET AL. (1993) und MUNISHI (1998) sowie die Schemata in WAHRIG (1997) sind hierfür exemplarisch. Ein funktionierendes Morphologiesystem ist Voraussetzung für eine maschinelle Syntaxanalyse, es sei denn, die Syntax verwendet ein Vollformlexikon, was für die form- und typreiche albanische Morphologie (und Wortbildung) noch weniger praktikabel wäre als für viele andere Sprachen.

(2) Eine korpusbasierte Satztypologie und Untersuchungen zur Valenz, am besten die Erstellung eines korpusbasierten Valenzwörterbuches.

---

[7] Siehe auch *http://www.lissus.com/albanian/index.htm* . Das Wörterbuch ist in ECI/MCI (1994) auch enthalten.

Diese sind für eine Modellierung der Syntax sehr wichtig. Gute Beispiele wären hier HELBIG/SCHENKEL (1991), HERBST ET AL. (2004) und SCHUMACHER ET AL. (2004).
(3) Die Erstellung lexikalisch-semantischer Wortnetze für das Albanische. Dies ist in Zukunft unumgänglich, um die Semantik des Albanischen behandeln zu können. *WordNet*[8] sowie Wörterbücher vom Typ „Wortschatz nach Sachgruppen" (DORNSEIFF 2004) könnten als Prototyp dienen.

5 Computerlinguistische Programme
Bedingt vor allem durch das Fehlen formalisierten und maschinenverwertbaren Wissens und einer albanischen Schule der Computerlinguistik beziehungsweise linguistischen Informatik, existieren bisher nur wenige Systeme bzw. Programme für das Albanische. Zu nennen wären hier: TROMMER (1997): Ein Modell für die albanische Verbflexion; KABASHI (1999): Ein Modell für die Hauptkonstruktionen der albanischen Syntax mit einer kleinen Morphologie; KABASHI (2003): Ein System zur morphologischen Analyse von Verben; und TROMMER und KALLULLI (2003): Ein morphologisches Taggingsystem.[9]
Gestützt auf das oben genannte System zur morphologischen Analyse des Albanischen wird zur Zeit von KABASHI (2007) ein System zur syntaktischen Analyse entwickelt, das sich insbesondere mit der Modellierung der Valenz und der klitischen Pronomina befasst.[10] Die dem System zugrunde liegende Morphologie wird momentan auch für die nominale Analyse erweitert.

6 Modellierung linguistischen Wissens
Zunächst sollte das linguistische Wissen möglichst korrekt, vollständig und detailliert erfasst werden, wobei auf eine adäquate Modellierung dieses Wissens zu achten ist. Letzteres ist notwendig, um eine computergestützte Verarbeitung zu gewährleisten, besonders, wenn die Wissenskomponente als Modul für große beziehungsweise verschiedene Projekte verwendet und wiederverwendet werden soll. Dies macht eine Abspeicherung in übersichtlicher, strukturierter und plattformunabhängiger Form, beispielsweise in XML (*Extensible Markup Language*), erstrebenswert. Oft scheitern Systeme gerade an einer schlecht durchdachten Modellierung. Schließlich spielt neben der Erschließung neuer Daten und

[8] *http://wordnet.princeton.edu*
[9] *http://www.iit.demokritos.gr/skel/bci03_workshop/papers/SESSION4_2-9_Trommer.pdf*
[10] *http://www.linguistik.uni-erlangen.de/~bmkabash/index.html*

der Modellierung neuen Wissens auch die Pflege und Aktualisierung der bestehenden Ressourcen eine bedeutende Rolle. Lexika beispielsweise sollten a priori so konzipiert und modelliert werden, dass aus einer Quelle verschiedene Formate generiert werden können, so dass neben herkömmlichen gedruckten auch maschinenlesbare (*machine readable*) und maschinenverarbeitbare (*machine tractable*) Versionen bereitgestellt werden können.

Literatur

[BUCHHOLZ ET AL. 1993] Buchholz, Oda / Wilfried Fiedler / Gerda Uhlisch: *Wörterbuch Albanisch–Deutsch*. München: Langenscheidt, Verlag Enzyklopädie, 1993. Erste Auflage Leipzig 1977.

[ÇABEJ 1987] Çabej, Eqrem: *Meshari i Gjon Buzukut (1555). I-II*. Prishtinë: Rilindja, 1987.

[DORNSEIFF 2004] Dornseiff, Franz: *Der deutsche Wortschatz nach Sachgruppen*. 8. Auflage. Hrsg. von Uwe Quasthoff. Mit einer lexikographisch-historischen Einführung und einer ausgewählten Bibliographie zur Lexikographie und Onomasiologie von Herbert Ernst Wiegand. Berlin / New York: Walter de Gruyter, 2004.

[DHRIMO ET AL. 2002] Dhrimo, Ali et al.: *Fjalor sinonimik i gjuhës shqipe*. Tiranë: Toena, 2002.

[ECI/MCI 1994] ELSNET: *European Corpus Initiative, Multilingual Corpus I (ECI/MCI) CD-ROM*. Utrecht: ELSNET, 1994.

[HELBIG/SCHENKEL 1991] Helbig, Gerhard / Wolfgang Schenkel: *Wörterbuch zur Valenz und Distribution deutscher Verben*. 8. Auflage. Tübingen: Max Niemeyer Verlag, 1991. Erste Auflage Leipzig 1969.

[HERBST ET AL. 2004] Herbst, Thomas / David Heath / Ian F. Roe / Dieter Götz: *A Valency Dictionary of English. A Corpus-Based Analysis of the Complementation Patterns of English Verbs, Nouns and Adjectives*. Berlin / New York: Mouton de Gruyter, 2004.

[KABASHI 1999] Kabashi, Besim: *Modellierung und Implementierung einer LA-Grammatik für ein Fragment des Albanischen*. Unveröffentlichtes Manuskript. Computerlinguistik, Universität Erlangen–Nürnberg, 1999.

[KABASHI 2003] Kabashi, Besim: *Automatische Wortformerkennung für das Albanische*. Magisterarbeit im Fach „Linguistische Informatik". Universität Erlangen–Nürnberg, 2003.

[KABASHI 2005] Kabashi, Besim: „Analiza automatike e fjalëformave të gjuhës shqipe". In: *Seminari Ndërkombëtar për Gjuhën, Letërsinë dhe Kulturën Shqiptare*, XXIII. Universiteti i Prishtinës. Prishtinë, 16-28.08.2004. Libri 23/1, 129-135.

[KABASHI 2006/Im Druck] Kabashi, Besim: „Disa propozime për modelimin e informacionit në leksikografinë kompjuterike". In: *Seminari Ndërkombëtar për Gjuhën, Letërsinë dhe Kulturën Shqiptare*, XXIV. Universiteti i Prishtinës. Prishtinë, 22.08-03.09.2005.

[KABASHI 2007/Im Druck] Kabashi, Besim: "Pronominal clitics and Valency in Albanian. A computational linguistics prespective and modelling within the LAG-Framework". In: Herbst, Thomas / Katrin Götz (Eds.): *Valenz / Valency. Theoretical, descriptive and cognitive issues*. Berlin: Mouton de Gruyter, 2007.

[FJDSH 1976] Kostallari, Androkli et al.: *Fjalori drejtshkrimor i gjuhës shqipe*. Tiranë: Akademia e Shkencave e RPS të Shqipërisë. Instituti i Gjuhësisë dhe i Letërsisë, 1976.

[FJALORI 1984] Kostallari, Androkli et al.: *Fjalor i shqipes së sotme*. Tiranë: Akademia e Shkencave e RPS të Shqiperisë. Instituti i Gjuhësisë dhe i Letërsisë, 1984.

[FJALORI 2002] Thomai, Jani et al. / Fatmir Toçi (Hg.): *Fjalor i shqipes së sotme*. Tiranë: Akademia e Shkencave e Shqipërisë. Instituti i Gjuhësisë dhe i Letërsisë & Toena, 2002.

[FJS 2005] Jani Thomai et al.: *Fjalor sinonimik i gjuhës shqipe*. Tiranë: Akademia e Shkencave e Shqipërisë. Instituti i Gjuhësisë dhe i Letërsisë, 2005.

[MUNISHI 1998] Munishi, Zijadin: *Zgjedhimi i foljeve*. Prishtinë: Libri Shkollor, 1998.

[RESSULI 1958] Ressuli, Namik: *Il »Messale« di Giovanni Buzuku*. Città del Vaticano: Biblioteca Apostolica Vaticana, 1958.

[SAMARA 1998] Samara, Miço: *Fjalor i antonimeve në gjuhën shqipe*. Shkup: Shkupi, 1998.

[SCHUMACHER ET AL. 2004] Schumacher, Helmut / Jacqueline Kubczak / Renate Schmidt / Vera de Ruiter: *VALBU – Valenzwörterbuch deutscher Verben*. Tübingen: Gunter Narr Verlag, 2004.

[SNOJ 1994] Snoj, Marko: *Rückläufiges Wörterbuch der albanischen Sprache*. Hamburg: Buske Verlag, 1994.

[TROMMER 1997] Trommer, Jochen: *Eine Theorie der albanischen Verbflexion in mo_lex*. Magisterarbeit. Universität Osnabrück, 1997

[TROMMER/KALLULLI 2003] Trommer, Jochen / Dalina Kallulli: "A Morphological Tagger for Standard Albanian". *Workshop on*

Balkan Language Resources and Tools, 21 November 2003, Thessaloniki, Greece.

[THOMAI 1999] Thomai, Jani: *Fjalor frazeologjik i gjuhës shqipe*. Tiranë: Akademia e Shkencave e Republikës së Shqiperisë. Instituti i Gjuhësisë dhe i Letërsisë & Shkenca, 1999.

[WAHRIG 1997] Wahrig, Gerhard: *Deutsches Wörterbuch*. 6. Auflage. Neu herausgegeben von Renate Wahrig-Burfeind. Güttersloh: Bertelsmann Lexikon Verlag, 1997.