ABSTRACTS

Arbeitsgruppe KORPUSBASIERTE LINGUISTIK

@ 40. Österreichische Linguistiktagung

22.-23. November 2013

Veranstaltet vom
ICLTT Institut für Corpuslinguistik und Texttechnologie
Kooperation mit
BaC:us

Andrea Abel, Aivars Glaznieks, Egon Stemle (Bozen)

Automatische Annotation von Schülertexten – Herausforderungen und Lösungsvorschläge am Beispiel des Projekts KoKo

Der Vortrag stellt den iterativen Workflow zur Erstellung eines lemmatisierten, POS-getaggten und nach ausgewählten sprachlichen Merkmalen annotierten Lernerkorpus vor und geht auf Schwierigkeiten und Besonderheiten bei der Korpuserstellung mit L1-Lernertexten ein.

Lernertexte weisen häufig Schreibweisen und Konstruktionen auf, die der Standardsprache nicht entsprechen. Da korpuslinguistische Verarbeitungstools gewöhnlich Zeitungstexte als Eingabe erwarten, können Lernertexte bei der automatischen Verarbeitung Schwierigkeiten bereiten. Dadurch kann die mitunter sehr hohe Zuverlässigkeit der Tools (z.B. eines POS-Taggers, Giesbrecht & Evert 2009) erheblich herabgesetzt. Eine Herausforderung bei der korpuslinguistischen Aufbereitung von Lernertexten liegt folglich darin, ihre Merkmale im Workflow so zu berücksichtigen, dass sie trotz der Abweichungen vom Standard mit einer ähnlichen Zuverlässigkeit verarbeitet werden können wie standardsprachliche Texte.

Im Projekt „KoKo“ wurden rund 1300 Schülertexte (811.330 Tokens) aus Oberschulen in Thüringen, Nordtirol und Südtirol für ein deutschsprachiges L1-Lernerkorpus aufbereitet. Mit o.g. Abweichungen wurde dabei folgendermaßen umgegangen: Bereits bei der Digitalisierung der handschriftlichen Daten wurden die Transkripte mit zusätzlichen Annotationen versehen, die Orthographiefehler, okkasionelle Kurzwortbildungen, Emotikons u.Ä. erfassen. Nachfolgend wurde das Korpus lemmatisiert und getaggt. In einem separaten Verarbeitungsschritt wurden mithilfe des POS-Taggers nicht automatisch verarbeitete Textmerkmale ermittelt, die anschließend entweder manuell annotiert oder dazu verwendet wurden, den Tagger neu zu trainieren. Der dadurch in Gang gesetzte iterative Prozess der Korpuserstellung ermöglicht es, die Qualität der Lemma- und POS-Annotationen des L1-Lernerkorpus sukzessiv zu verbessern. Diese iterative Herangehensweise kann auch für die mögliche Annotation weiterer Ebenen beibehalten werden (vgl. Voormann & Gut 2008).

Literatur:

Giesbrecht, Eugenie & Evert, Stefan (2009): Is Part-of-Speech Tagging a Solved Task? In: Proceedings of the Fifth Web as Corpus Workshop (WAC5), 27-35. San Sebastián, Spain.

Voormann, Holger & Gut, Ulrike (2008): Agile corpus creation. In: Corpus Linguistics and Linguistic Theory 4-2, 235-251.

Peter Ďurčo (Trnava, Bratislava)

Korpusmorphologie als theoretische und empirische Basis für die Kollokationslexikographie

Das zentrale theoretische Problem ist die Delimitation einer Wortverbindung als einer freien oder usualisierten und damit beschreibungswürdigen Wortschatzeinheit. Die korpusbasierten Analysen bieten das Instrumentarium, diese Usualität von Wortverbindungen völlig neu quantitativ zu verifizieren. Aus dem Korpus wird eine möglichst große Anzahl sprachlicher Einheiten generiert bzw. ihr Usualitätsgrad mittels Korpusdaten überprüft und dann als Referenzinventar zur Verfügung gestellt.

Das Konzept basiert auf der Idee einer „Corpusmorphologie". Man geht hier von der Hypothese aus, dass die typischen, usuellen, lexikalisierten und idiomatisierten Wortkombinationen primär nicht auf der Wortebene, sondern auf der Wortformebene entstehen. Durch die schrittweise Emanzipation und Separation dieser Kombinationen und durch Veränderungen in ihren regulären paradigmatischen und syntagmatischen Eigenschaften kommt es zu ihrer neuen Funktionalität. Auf Grund von verschiedenen Testverfahren (wir schlagen die sog. 4K-Methode vor: Kollokationstest, Kategorientest, Kommutationstest und Kompositionalitätstest) kann man eine Art Klassifikation und Typologie der festen Wortverbindungen mit relativ exakten Prozeduren festlegen.

Die korpusbasierte Aufwertung von Sprachdaten als zentrale lexikografische Information führt auch zu neuen Produkttypen (z.B. kann man Taxonomien usueller Wortverbindungen durchaus schon als relevante lexikografische Produkte ansehen). Die Inventare können dazu dienen, abgesicherte Kandidaten für Wörterbücher oder verlässliche Kandidaten für didaktische Zwecke zu liefern.

Erste Ergebnisse in diesem Bereich bietet die WICOL-Plattform. Sie zeigt, wie die exhaustive Erarbeitung einer möglichst großen Menge von Kollokationen und Wortverbindungen (ein- und zweisprachig) auf der Basis einer komplexen Korpusmethodik realisiert werden kann.

Für jede Wortart wurde ein universelles kombinatorisches Modell in der Matrixform erarbeitet, das alle theoretisch denkbaren binären Kombinationen des Basiswortes mit anderen autosemantischen Wortarten voraussetzt. Die erstellten Kollokationsprofile stellen bei jedem Lexem das individuelle spezifische und einmalige Kollokationsparadigma des Wortes. Die Kollokationsprofile bilden die Basis für eine viel detailliertere lexikographische Beschreibung der Kollokabilität der Lexeme und ihrer Semantik.

Peter Ernst (Wien)

Eigennamen in der Korpuslinguistik

Obwohl die Computerlinguistik so alt wie die Computertechnologie ist, werden durch das Anwachsen von Speicher- und Zugangsmöglichkeiten erst in den letzten Jahren vermehrt brauchbare elektronische Korpora aufgebaut. Trotz zahlreicher und intensiver Versuche, einen einheitlichen Standard für ihre Erstellung zu erreichen, ist dies bis heute nicht gelungen und wird wohl auch in der nahen Zukunft nicht erreichbar sein.

Der Vortrag geht der Frage nach, ob und wie Eigennamen in verschiedenen Systemen abgebildet werden, und nähert sich dem Problem sowohl von der Input als auch der Output-Seite (Eingabe und Abfrage). Es werden verschiedene Korpora und ihre Lösungsansätze miteinander verglichen, darunter das Referenzkorpus des IDS Mannheim und das Wortschatzportal der Universität Leipzig.

Susanne Haaf (Berlin)

Strukturelle und linguistische Annotation in historischen Textkorpora am Beispiel des Deutschen Textarchivs

Innerhalb des BMBF-geförderten Infrastrukturprojekts CLARIN-D (http://www.clarin-d.de) werden Sprachressourcen verschiedener Anbieter gebündelt und gemeinsam verfügbar und durchsuchbar gemacht. In diesem Zusammenhang werden u. a. Best practices für Annotationsformate und -inhalte für die Auszeichnung von Textkorpora sowohl auf der Ebene der Transkription als auch auf der Ebene der Metadatenerfassung erarbeitet.

Das DFG-geförderte Projekt Deutsches Textarchiv (DTA) konzentriert sich als Partner im CLARIN-D-Verbund auf die Erstellung und Aufbereitung historischer Textressourcen (http://www.deutschestextarchiv.de). Dabei geht es zunächst um die Anwendung, Pflege und Weiterentwicklung des TEI/P5-basierten DTA-Basisformats (DTABf, vgl. http://www.deutschestextarchiv.de/doku/basisformat), welches das Best practice-Format für die strukturelle Aufbereitung historischer gedruckter Texte in CLARIN-D darstellt (vgl. http://www.clarin-d.de/de/sprachressourcen/benutzerhandbuch.html, Kap. II,6.) Volltexte historischer Quellen, die einerseits durch Eigendigitalisierung des DTA entstehen, andererseits im Rahmen externer Projekte erarbeitet wurden und in das DTA integriert werden, werden in das DTABf überführt, um eine einheitliche Strukturierung des gesamten DTA Korpus (DTA Kernkorpus + DTA Erweiterungskorpora) zu gewährleisten. Diese strukturelle Textauszeichnung entsprechend dem DTABf erfolgt inline, d. h. direkt in den Textdaten.

Sämtliche Texte des DTA durchlaufen sodann eine automatische linguistische Analyse, welche die Tokenisierung, Lemmatisierung, POS Analyse und orthographische Normalisierung der historischen Wortformen umfasst. Diese Informationen werden den jeweiligen Texten als stand-off-Annotation beigegeben, die in einem zweiten Schritt in das CLARIN-D-Format für die Textauszeichnung (Text Corpus Format; TCF) überführt wird.

Die Erfassung ausführlicher Metadaten zu sämtlichen DTA-Textressourcen und deren Quellen erfolgt ebenfalls entsprechend den Richtlinien des DTABf. Die TEI/P5-basierten DTABf-Metadaten können dann in das CLARIN-D-Metadatenformat CMDI sowie nach Dublin Core konvertiert werden.

Der Beitrag gibt Einblicke in die im DTA etablierte Praxis der linguistischen und strukturellen Textauszeichnung für historische Textkorpora und die Möglichkeiten, die sich daraus für die Korpusrecherche ergeben. Darüber hinaus wird gezeigt, wie durch die Konvertierungsmöglichkeiten des DTABf in andere Metadaten- und Textkorpusformate die Verfügbarkeit der DTA-Daten in anderen Kontexten, insbesondere innerhalb von CLARIN-D (z. B. über das Virtual Language Oberservatory VLO http://www.clarin.eu/vlo sowie die Federated Content Search http://weblicht.sfs.uni-tuebingen.de/Aggregator/), gewährleistet wird.

Miroslava Hliničanová, Matej Ďurčo, Karlheinz Mörth, Wolfgang U. Dressler

Phonotaktische versus morphonotaktische Konsonantengruppen im Slowakischen und Deutschen: Eine kontrastive corpuslinguistische Untersuchung

Am ICLTT werden in Zusammenarbeit mit der Arbeitsgruppe „Komparative Psycholinguistik“ des Instituts für Sprachwissenschaft der Universität Wien Eigenschaften von Konsonantengruppen untersucht, die entweder morphonotaktisch sind, d.h. durch morphologische Operationen entstehen, wie im Wortauslaut in D. *lach+st, höch+st* oder allgemein phonotaktischer Natur sind, wie in *Post, Rast.* Während die erste Konsonantengruppe nur durch morphologische Operationen

entsteht und die zweite im Wortauslaut nach Kurzvokal nie, gibt es Zwischenstufen, wie auslautendes /pst/, welches als default morphonotaktisch ist, z.B. in *kapp+st, lieb+st,* während nur wenige Wörter wie *Obst, Papst* phonotaktisch sind. Dieser morphonotaktische Default ist stärker als bei wortauslautendem /rst/, vgl. *irr+st,lehr+st* usw. vs. *erst, Durst* usw. (Dressler & Dziubalska-Kołaczyk 2006).

In unserem Beitrag werden wir kontrastiv-typologisch die phonotaktischen und morphonotaktischen Obstruentengruppen des Slowakischen und Deutschen im Anlaut, Inlaut und Auslaut vergleichen, gestützt auf die elektronischen Corpora beider Sprachen: Slovenský národný korpus, Austrian Media Corpus.

Im Wortanlaut ist das Slowakische viel reicher an Obstruentengruppen als das Deutsche, einerseits morphonotaktisch durch die Präfixe *z-, s-, v-,* z.B. in *s+chrad+nú-t´* ‚altern', *z+drav+i+t´* ‚begrüßen', *v+kvap+k+a+t´* ‚eintropfen', andererseits durch größere phonotaktische Komplexität, z.B. *pstruh* ‚Forelle', *škvár+a* ‚Schlacke', *vdov+a* ‚Witwe'. Der deutsche Standard hat keine monokonsonantischen Präfixe, im Gegensatz zu bair.-öst. *g'storben* usw. , und das Maximum an phonotaktischer Komplexität wird durch 2 Obstruenten plus Sonorant erreicht, wie in *Strafe, Sklave.*

Im Wortinlaut haben beide Sprachen vergleichbare phonotaktische Konsonantengruppen. Morphonotaktische Obstruentengruppen entstehen in der Wortbildung entsprechend dem typologischen Unterschied zwischen den germanischen kompositionsreichen Sprachen und den stärker derivationellen slawischen Sprachen. So haben deutsche Komposita wie *Dienst+geber* keine Entsprechungen im Slowakischen, nicht nur weil die slowakische Alltagssprache viel kompositionsärmer ist als die deutsche, sondern auch weil für die slowakische Komposition Interfigierung von *–o-* typisch ist, wie in *štrk+o+piesky* ‚Schotterstrand'. Hingegen hat das Deutsche sogar ein Obstruenteninterfix *–s-,* wie in *König+s+krone*. Die slowakische Wortbildung schafft durch viele konsonantenauslautende Präfixe und konsonantenanlautende Suffixe zahlreiche komplexe morphonotaktische Obstruentengruppen, wie *roz+drob+en+ý* ‚zerkleinert', *gróf+stvo* ‚Grafschaft'. Dazu kommt, dass die slowakische, aber nicht die deutsche Deklination, durch Vokaltilgung neue morphonotaktische Obstruentengruppen schafft, wie in *mozog* ‚Gehirn', Gen.Sg. *mozg+u*, *líst+ok* ‚Blättchen', Gen.Sg. *líst+k+a/u* vs. *po+tok* ‚Bach',Gen.Sg. *po+tok+a/u, otec* ‚Vater', Gen,Sg. *otc+a* . Im Wortauslaut kennt das Slowakische zum Unterschied vom Deutschen (s.o.) keine konsonantischen Flexionssuffixe, schafft aber im suffixlosen Genitiv Plural durch Tilgung des Stammvokals morphonotaktische Obstruentengruppen, wie in *rod+i+sk+o* ‚Geburtsort', Gen.Pl. *rodísk*, *mzd+a* ‚Lohn', Gen.Pl. *miezd*. Phonotaktische Auslautgruppen kennen beide Sprachen, sie sind aber im Deutschen weitaus häufiger.

Diese Unterschiede werden korpuslinguistisch durch Typen- und Tokenfrequenzen ergänzt, wobei die Frage der Unterscheidung starker und schwacher defaults problematisch ist. Zum Abschluss werden weitere Probleme der korpuslinguistischen Analyse besprochen.

Bernhard Hurch, Jennifer Brunner, Anneliese Kelterer (Graz)

Zur Struktur eines morphologisch basierten Wörterbuchs des Pame in Lexus

Das Pame, eine Oto-Mangue Sprache Zentralmexikos, ist schlecht, das *pame central*, mit dem wir uns in einem Grazer Dokumentationsprojekt beschäftigen, so gut wie gar nicht beschrieben. Die Untergruppe Oto-Pame ist für besonders komplexe morphologische Systeme bekannt, in denen alle verschiedenen formalen Verfahren parallel verwendet werden: Präfigierung (verschiedene Slots), Suffigierung (verschiedene Slots), Anlautsmutationen des Stammes, Vokalalternation des Stammes und Stammallomorphie.

Ziel des auf dem Workshop vorzustellenden Wörterbuchs ist es, eine Plattform zu schaffen, die die Verbindung von "Lexikon" und morphologischer Form und ev. morphologischer Ordnung

(Klassen?) erkennbar macht, gewissermaßen also ein Wurzelwörterbuch mit grammatischer Ausrichtung. Im Pame gibt es wahrscheinlich keine andere (lexikalische?, syntaktische?) primäre Wortklassenunterscheidung als die zwischen Funktionswörtern, Adverbialen und Inhaltswörtern, wobei letztere selbständig Prädikationen bilden. Unflektierte (infinite) Wortformen gibt es nicht, das Flexionspotential von Inhaltswörtern ist relativ ähnlich, egal ob es sich um stärker "nominale" oder "verbale" Wörter handelt. Wenn Lexeme Einheiten sind, die aus dem Inventar der Flexionsformen abstrahiert werden, läge es für das Pame nahe, die 3. Person (poss oder pers) als die am wenigsten markierte anzusetzen (ein Verfahren, das auch in exzellenten lexikographischen Arbeiten zu anderen mesoamerikanischen Sprachen verfolgt wird, üblich im Maya), doch brächte das den Nachteil, daß mit 3pers/poss sg/pl jeweils die Flexionsformen mit der stärksten Stammallomorphie Ausgangsformen wären. Das Pame besitzt die traditionellen europäischen nominalen Kategorien Kasus und Genus nicht, die Personenbezeichnungen sind für alle Inhaltswörter von primärer Bedeutung, ihre externe Markierung für alle Inhaltswörter ident (nicht 'homophon' wie Türkisch oder Ungarisch), doch sind diese Relationen primär flexivisch markiert, ebenfalls in starker Parallelität quer zu möglichen lexikalischen Kategorien. Die kurze Präsentation wird darauf abzielen, die morphologische Struktur direkt mit Wurzeln in Verbindung zu setzen und eventuelle Wortklassen als sekundäre Kategorisierungsmöglichkeit zu präsentieren. Wir werden ebenfalls die Wahl von lexus als Plattform rechtfertigen.

Katharina Korecky-Kröll (Wien)

Semi-automatische lexikonbasierte morphologische Kodierung und Analyse mit CHILDES: Erfahrungen mit kindersprachlichen Spontansprachkorpora und erste Arbeiten zu einem rein erwachsenensprachlichen Spontansprachkorpus

Das System CHILDES (Child Language Data Exchange System, 1984 von Brian MacWhinney und Catherine Snow begründet) ist die für die Kindersprachforschung entwickelte Komponente des TalkBank-Systems (http://www.talkbank.org) und eine der führenden Methodologien in diesem Bereich. CHILDES besteht aus drei Hauptkomponenten:

1) CHAT (Codes for the Human Analysis of Transcripts): standardisiertes Format für die Transkription von Kindersprachdaten
2) CLAN (Computerized Language Analysis): Programmpaket, das die mehr oder weniger automatisierte Analyse von Daten im CHAT-Format ermöglicht.
3) CHILDES database: Datenbank, aus der im CHAT-Format verschriftete Kindersprachdaten in derzeit 34 verschiedenen Sprachen für Forschungszwecke verwendet werden können.

Unsere Arbeitsgruppe verfügt über drei umfangreiche Korpora von Wiener Kindern, die über mehrere Jahre in spontaner Interaktion mit ihren Hauptbezugspersonen aufgenommen wurden. Diese Daten wurden im CHAT-Format transkribiert und kodiert und bereits unter vielen Gesichtspunkten analysiert.

Derzeit werden im Rahmen des vom Wiener Wissenschafts-, Forschungs- und Technologiefonds geförderten Projekts „INPUT – Investigating Parental and Other Caretakers' Utterances to Kindergarten Children" weitere Spontansprachdaten von insgesamt 61 Wiener Kindern und ihren Hauptbezugspersonen gesammelt, verschriftet, kodiert und analysiert. Anhand von Beispielen aus diesem Projekt wird das Transkriptionssystem vorgestellt und die semi-automatische lexikonbasierte morphologische Kodierung erläutert. Es folgt eine kurze Erklärung der wichtigsten Basisanalysen zu Lemma-, Type- und Tokenfrequenzen, mittlerer Äußerungslänge (MLU) und lexikalischer Diversität (VOCD), die mit dem CLAN-Programmpaket automatisch durchgeführt werden können.

Seit Sommer 2013 liegen uns auch erste Daten (zwei transkribierte und kodierte Aufnahmestunden) aus einem insgesamt 40 Stunden umfassenden Spontansprachkorpus von Wiener Erwachsenen vor. Allererste Vergleiche zwischen kindgerichteter Sprache (CDS) und erwachsenengerichteter Sprache (ADS) zeigen große Unterschiede im Wortschatz und gewisse Unterschiede in einigen Aspekten der Morphologie.

Literatur:

1) *Dressler, W.U., K. Korecky-Kröll, C. Czinglar & K. Uzunkaya-Sharma. eingereicht. Caretaker input to, and output of, bilingual children at home and in kindergarten: Filling a European lacuna in the causal chain leading to dispriviledged language competences. Festschrift for Alberto Mioni.*

2) *Korecky-Kröll, K. 2011. Der Erwerb der Nominalmorphologie bei zwei Wiener Kindern: Eine Untersuchung im Rahmen der Natürlichkeitstheorie. Universität Wien: Dissertation.*

3) *Laaha, S. & K. Korecky-Kröll. im Druck: Verschriftung, Kodierung und Analyse mit CHILDES. Erscheint in: E. Wandl-Vogt & K. Korecky-Kröll. Hrsg. Transkriptionssysteme im Vergleich: Sprache - Ton - Bild. Kodierung gesprochener Sprache. Wien: Präsens.*

4) *MacWhinney, B. 2000. The CHILDES Project: Tools for Analyzing Talk. 3rd edition. Mahwah, NJ: Erlbaum.*

Karlheinz Mörth (Wien)

CLARIN.AT: Digitale Infrastrukturen für die Linguistik

Digitale Sprachressourcen spielen in vielen Bereichen linguistischen Arbeitens eine zunehmend bedeutende Rolle, dies trifft insbesondere auf Ansätze zu, in denen mit empirischen Daten gearbeitet wird. Neben Sprachdaten wie digitalen Wörterbüchern, digitalen Corpora, Datenbanken etc. geht es dabei auch um digitale Werkzeuge, Dokumentationen, Standards und *Best practices*, die die Wiederverwertbarkeit derartiger Ressourcen überhaupt erst ermöglichen.

In Europa beteiligt sich eine Reihe transnationaler Initiativen seit geraumer Zeit am Aufbau von digitalen Forschungsinfrastrukturen. Von besonderer Bedeutung für SprachwissenschaftlerInnen sind unter diesen besonders DARIAH (Digital Research Infrastructure for the Arts and Humanities) und CLARIN (Common Language Resources and Technology Infrastructure). Während DARIAH sich an alle GeisteswissenschaftlerInnen wendet, verfolgt CLARIN enger definierte, primär an Sprache orientierte Interessen. CLARIN hat von der Kommission der Europäischen Union im vergangenen Jahr den offiziellen Status eines *European Research Infrastructure Consortiums* (ERIC) erhalten und bemüht sich darum, WissenschaftlerInnen in den Sozial-, Kultur- und Geisteswissenschaften den einfachen und nachhaltigen Zugang zu digitalen Sprachdaten und fortgeschrittenen Werkzeugen zur Auffindung, Erforschung, Annotation, Analyse oder Kombination solcher Daten zu ermöglichen.

In beiden Infrastrukturen haben sich österreichische ForscherInnen von allem Anfang an engagiert. Österreich ist Gründungsmitglied sowohl von CLARIN als auch von DARIAH. Die Präsentation wird auf rezente, insbesondere für die Linguistik relevante Entwicklungen eingehen, kurz über bereits Erreichtes berichten und einen Ausblick auf geplante Vorhaben geben.

Friedrich Neubarth, Tina Hildenbrandt, Sylvia Moosmüller (Wien)

Korpusbasierte maschinelle Übersetzung vom Standarddeutschen in den Wiener Dialekt

Üblicherweise werden für die statistische maschinelle Übersetzung (SMT) große Mengen an bilingualen Textdaten benötigt, um hinreichend brauchbare statistische Modelle erstellen zu können. Diese sind im Kontext von dialektalen Varietäten gewöhnlich nicht gegeben. Allerdings

stehen dialektale Varietäten zu Standardsprachen mit umfangreichen Ressourcen in einem engen linguistischen Naheverhältnis, sodaß es möglich sein sollte, diese Nähe oder Ähnlichkeit für die Zwecke der Datengenerierung gewinnbringend einzusetzen.

In einem aktuellen Projekt wird ein SMT System für das Sprachpaar Deutscher Standard – Wiener Dialekt entwickelt, wobei der Umgang mit dem Mangel an bilingualen Daten methodisch im Vordergrund steht. Der Ansatz verfolgt naturgemäß eine 'bootstrapping' Strategie: zu Beginn wurde ein relativ kleines Korpus an bilingualen Sätzen (ca. 6700) aus Transkripten in einer eigens entwickelten Orthographie von (möglichst authentischen) Sprachdaten des Wiener Dialekts erstellt. Die daraus entwickelten SMT Modelle dienen der Selektion von informationsreichen, relevanten Sätzen ('active learning'), die in der Folge dem Korpus hinzugefügt werden. Zusätzlich sind wir bestrebt, online verfügbare parallele Ressourcen zu verwerten. Zwischen der bairischen und der deutschen Wikipedia existieren über 5000 parallele Artikel. 202 davon sind explizit als Wiener Dialekt ausgewiesen, aus diesen konnten wir weitere 4414 parallele Satzpaare extrahieren. Der Algorithmus zur Identifikation solcher Satzpaare basiert auf der relativen Ähnlichkeit zwischen bairischen Dialekten und der Standardvarietät. Eine Bewertungsfunktion weist jedem Satzpaar einen Ähnlichkeitswert zu, ausgehend von dem höchsten Wert werden die Sätze als zugehörig markiert, bis ein gewisser Schwellwert erreicht ist. Ein Problem, das noch der Lösung harrt, ist dabei die Notwendigkeit, die vielfältigen orthographischen Varianten auf eine einzige, konsistente Variante zu normalisieren, um die Anwendbarkeit von maschinellen Lernmethoden zu gewährleisten.

Das SMT System, das in dem Projekt zur Anwendung kommt, macht sich implizit auch die linguistische Nähe der zwei Varietäten zunutze: auf der höheren Wort-Ebene kommt ein phrasenbasiertes Modell zum Einsatz, das zwar nicht besonders groß, aber in der Lage ist, typische Phrasen oder Lexeme des Dialektes zu inkorporieren. Für die Behandlung der zahlreichen 'out-of-vocabulary' Wörter wird ein Modell auf Zeichenebene trainiert, das, obgleich weniger akkurat, durch das Erlernen von korrespondierenden Zeichenketten einen viel weiteren Skopus hat. Die bis dato besten Ergebnisse wurden durch eine Kombination dieser zwei Ebenen erzielt.

Jutta Ransmayr / Matej Ďurčo / Karlheinz Mörth

AMC - Austrian Media Corpus: Korpusbasierte Forschungen zum österreichischen Deutsch

Das Institut für Corpuslinguistik und Texttechnologie (ICLTT) der österreichischen Akademie der Wissenschaften hat eine neue Sprachdatenbank aufgebaut, die bisherige Maßstäbe sprengt – das „Austrian Media Corpus" (AMC). Die Datenbank umfasst rund 8 Milliarden running words, damit ist das AMC derzeit das größte Text-Corpus im deutschen Sprachraum.

Ermöglicht wurde dies durch eine enge Zusammenarbeit des ICLTT mit der Austria Presse Agentur (APA), die dem ICLTT große Teile ihrer digitalen Archivbestände für wissenschaftliche Zwecke zur Verfügung stellt. Dazu gehören alle digital verfügbaren APA-Pressemeldungen seit 1955, fast alle Tages- und Wochenzeitungen und die wichtigsten Magazine Österreichs seit Beginn der 1990er Jahre sowie Transkripte österreichischer TV-Nachrichtensendungen.

Durch die enorme Bandbreite der Quelltexte kommt ein wertvoller Sprachdatenpool zustande, der mehrere Jahrzehnte an österreichischen Texten umfasst. Durch diese Zusammensetzung der Datensammlung haben die Forscher der ÖAW nun Forschungsmaterial zum Sprachgebrauch in Österreich über die letzten Jahrzehnte, das im deutschsprachigen Raum sowohl qualitativ als auch quantitativ einzigartig ist.

Das „AMC" wird bereits in mehreren laufenden, lexikographisch orientierten Projekten eingesetzt, beispielsweise in der computerlinguistischen Unterstützung der bevorstehenden

Neuauflage des Österreichischen Wörterbuchs. Für das Wörterbuchprojekt „Variantenwörterbuch des Deutschen NEU“ lieferte das ICLTT im Rahmen einer Kooperation mit dem Institut für Germanistik der Universität Wien eine umfassende Datenbasis zum Sprachraum Österreich. Und auch das offizielle Österreich kann vom AMC profitieren: Im Rat für deutsche Rechtschreibung etwa können bei länderspezifischen Rechtschreibanalysen nun auch erstmals Auswertungen zum österreichischen Orthographiegebrauch Eingang in die Diskussion finden und für Empfehlungen des Rats berücksichtigt werden.

Der vorliegende Beitrag präsentiert das „Austrian Media Corpus“ und diskutiert exemplarisch Anwendungsbeispiele sowie die damit verbundenen technischen Herausforderungen und Lösungen im Umgang mit solch großen Datenmengen.

Claudia Resch / Ulrike Czeitschner / Karlheinz Mörth / Barbara Krautgartner / Eva Wohlfarter (Wien)

ABaC:us für LinguistInnen: Morphosyntaktische Annotation im „Austrian Baroque Corpus“

Das „**A**ustrian **Ba**roque **C**orp**us**“ ist eine strukturierte digitale Sammlung von Texten aus der Barockzeit, die am Institut für Corpuslinguistik und Texttechnologie aufgebaut und beforscht wird.

Um mit dieser digitalen Ressource neben – literatur- und kulturwissenschaftlichen Fragestellungen – auch und vor allem linguistische Forschungsfragen beantworten zu können, wurde bislang ein Großteil von ABaC:us (ca 180.000 Token) mit morpho-syntaktischen Annotationen (PoS-tagging und Lemmatisierung) versehen. Die Probleme, auf die man bei der Annotation von historischen, von der heutigen Norm abweichenden Sprachdaten stößt, sind hinlänglich bekannt: Der Vortrag geht daher auf Standards und Tools ein, die nutzbringend zur automatischen Annotation des Datenmaterials einerseits und zur regelbasierten manuellen Nachkorrektur der generierten Resultate andererseits eingesetzt wurden.

Das Ergebnis dieses im Rahmen des Projekts erprobten Workflows ist ein Gold-Standard-Korpus, dessen Datenbasis nun für die Annotatiion weiterer Texte herangezogen werden kann. Mit Evaluierungsverfahren anhand größerer Datenmengen können wir mittlerweile einschätzen, wie deutlich sich die automatisch generierten Ergebnisse durch die erarbeitete Datenbasis verbessern lassen. Abschließend werden Nutzungsperspektiven der linguistischen Abfrage und Pläne zur Weiterverwendung der historischen Korpusdaten skizziert.

Janusz Taborek (Poznań)

Korpusbasierte kontrastive Grammatik und Parallelkorpora

Das Ziel des Beitrags ist es, den Einsatz der mono- und bilingualen Korpora in der kontrastiven deutsch-polnischen (Lexiko-)Grammatik zu präsentieren. Die Entwicklung der korpusbasierten kontrastiven Linguistik, von Johansson für das Sprachenpaar Norwegisch-Englisch eingeleitet und von u.a. Aijmer/Altenberg (2013) als „a new era in contrastive linguistics“ bezeichnet, wird anhand der satzeinbettenden Prädikate im Deutschen und Polnischen dargestellt – aus der Perspektive des Verbs und aus der Perspektive des vom Verb (Prädikat) selegierten Nebensatzes.

Im ersten Teil des Beitrags werden satzförmig realisierte Subjekte (Subjektsätze) und die Klassifizierung der diese Subjektsätze einbettenden Prädikate dargestellt. Die Analyse der Verben erfolgt mithilfe der monolingualen Korpora (DeReKo und IPI PAN) für beide Sprachen (vgl. Taborek 2008, Engelberg/Cosma 2014).

Im zweiten Teil des Beitrags werden Ergebnisse einer korpus-basierten Studie des *dass*-Satzes im Hinblick auf (i) seine syntaktische Funktion, (ii) das einbettende Prädikat und (iii) die Äquivalente in der polnischen Sprache präsentiert. Diese Analyse erfolgt mithilfe des multilingualen Korpus der slawischen Sprachen PARASOL am Beispiel der literarischen Sprache.

Literatur:

Aijmer, K. und Altenberg, B. (2013). Advances in Corpus-based Contrastive Linguistics. Studies in honour of Stig Johansson. Amsterdam/Philadelphia.

Engelberg, S. und Cosma, R. (2014). „Subjektsätze als alternative Valenzen im Deutschen und Rumänischen." In Cosma, R., Engelberg, S., Schlotthauer, S., Stanescu, S. und Zifonun, G. (Hrsg.), Komplexe Prädikationen als Argumente. Kontrastive Untersuchungen zum Deutschen, Rumänischen und Englischen, Berlin (im Druck).

Taborek, J. (2008). Subjektsätze im Deutschen und im Polnischen. Syntaktisches Lexikon Subklassifizierung der Verben. Frankfurt a.M. etc.

Taborek, J. (2013). „Der dass-Satz im Deutschen und seine polnischen Äquivalente. Eine korpusbasierte Übersetzungsanalyse am Beispiel des Romans Parfüm von Patrick Süskind." (im Druck).

Branko Tošović (Graz)

Die morphosyntaktische Annotation im Gralis-Korpus

An der Karl-Franzens-Universität Graz wurde ein komplexes und mehrsprachiges Korpus der geschriebenen und gesprochenen Sprache entwickelt (http://www-gewi.kfunigraz.ac.at/gralis/index.html), das für Analysen und das Erlernen aller slawischen Sprachen und von deren Paralleltexten im Deutschen dient. 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 verwandten (slawischen) oder nicht bzw. weniger verwandten Sprachen (slawisch-deutsch) liegt.

1. Das Korpus besteht aus zwei Subsystemen – einem audiellen 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. Das Fix-Korpus umfasst Aufnahmen eines vorgegebenen Textes. Das Frei-Korpus setzt sich aus Aufnahmen spontaner Rede zusammen.

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

2. Typologisch gehört das Gralis-Korpus zu den lemmatisierten Korpora. Bislang ist eine morphosyntaktische Annotierung und die Suche nach entsprechenden Annotierungen für die Sprachen Serbisch, Kroatisch, Bosni(aki)sch und Montenegrinisch möglich.

3. Die morphosyntaktische Annotierung und die Suchabfrage erfolgen mit dem Gralis-MorphoGenerator. Er dient für (1) die automatische morphosyntaktische Annotierung der Wörter im mehrsprachigen Gralis-Korpus, (2) die morphologische Analyse der serbischen, kroatischen und bosnischen/bosniakischen Sprache, (3) die automatische Generierung der Paradigma aller

flektierten Lexeme. Die Annotation wurde im Jahr 2008 begonnen und 2013 beendet. Bis heute wurden 108.802 Lexeme morphologisch annotiert.

Martina Werner (Stuttgart) / Claudia Resch (Wien)

Paradigmatisierung von Fugenelementen anhand des Austrian Baroque Corpus (ABaC:us)

Die Musterbildung (Paradigmatisierung) von Fugenelementen in deutschen Komposita ist eine der intensiv diskutierten Fragen der (historischen) Morphologie des Deutschen. Im Gegenwartsdeutschen gelten insbesondere abgeleitete Substantive als interfigierungsaffin (wie etwa *Lehrling-s-vertrag, Altertum-s-forscher, Schönheit-s-operation*. Unter der Trennung von produktiven (i.S.v. reihenbildenden) und inproduktiven Suffixen (sog. ‚Genusmarker' nach WEGENER 2000) kommt synchron dem Genus femininum, nicht aber den anderen beiden Genera, eine zentrale Rolle bei der Interfigierung zu (sog. unparadigmatische Interfigierung wie in *Umgehung-s-straße*). Der Vortrag geht der Frage nach, welche ersten Muster der Interfigierung von suffigierten Substantiven in Erstposition fnhd. und früher nhd. Komposita sich beobachten lassen, um die Frage zu klären, ob sich – neben den synchronen Befunden – auch diachron eine ‚Genusinterfigierungspräferenz' in statu nascendi der Interfigierung nachweisen lässt. Gleichzeitig soll beleuchtet werden, wie die „unparadigmatische" Interfigierung der Feminina zustande kam, deren Existenz grammatiktheoretisch paradox ist.

Die empirischen Belege stammen aus dem Austrian Baroque Corpus (ABaC:us), das derzeit am Institut für Korpuslinguistik und Texttechnologie aufgebaut und beforscht wird.

Martina Werner (Stuttgart/Wien) / Karlheinz Mörth / Wolfgang U. Dressler (Wien)

Pluraldubletten im Deutschen

Empirisch ausgehend von der Dokumentation von Pluralvarianten im Grimmschen Wörterbuch soll die Frage im Vordergrund stehen, ob und inwieweit die historisch bezeugten Varianten auch im Neuhochdeutschen, speziell auch im Österreichischen Deutsch, noch bezeugt sind. Hierzu wurde das ICLTT-interne Austrian Academy Corpus sowie das Austrian Media Corpus beforscht. Die Ergebnisse der corpusbasierten Untersuchung sollen mit aktuellen Ergebnissen zur Pluralallomorphie des Deutschen kontrastiert werden, insbesondere vor dem Hintergrund der Frage, ob sich daraus Entwicklungstendenzen innerhalb der Deklinationsparadigmen des Deutschen ableiten lassen, die auf eine Re- oder Umorganisation des nominalen Flexionssystems des Deutschen hinweisen.

Theoretischer Ausgangspunkt wird eine Einleitung zu den morphologietheoretischen Konzepten overabundance, Heteroklisie, Plural und Pluraldubletten sein. Dann wird begründet, warum wir uns aus Zeitgründen auf die spezifische Auswahl von Pluraldubletten beschränken, die entweder einen *-s*-Plural enthält oder einen *-(e)n*-Plural mit heute illegalem Umlaut bzw. einen*–e*-Plural mit und ohne Umlaut. Ausgeschlossen wird Variation mit Nullplural sowie bei inzwischen ausgestorbenen oder homophonen Lemmata. Zum Abschluss werden Grenzen, Möglichkeiten und Perspektiven corpuslinguistischer Verfahren besprochen.