DOI 10.1515/zgl-2014-0004 — ZGL 2014; 42(1): 100–110 DE GRUYTER

Bericht

Monika Eller, Hagen Hirschmann

Modellierung nichtstandardisierter Schriftlichkeit

AG im Rahmen der 35. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft (DGfS), Universität Potsdam vom 12. bis 15. März 2013

Monika Eller: Ruprecht-Karls-Universität Heidelberg, Anglistisches Seminar, Kettengasse 12, D-69117 Heidelberg, E-Mail: monika.eller@as.uni-heidelberg.de
Hagen Hirschmann: Humboldt-Universität zu Berlin, Institut für deutsche Sprache und Linguistik, Unter den Linden 6, D-10099 Berlin, E-Mail: hirschhx@hu-berlin.de

Die nunmehr 35. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft (DGfS) wurde vom 12. bis 15.03.2013 von der Universität Potsdam ausgerichtet. Mit ihren 14 Arbeitsgruppen deckte die unter dem Rahmenthema *Informationsstruktur* stehende Tagung auch im Jahr 2013 wieder ein breitgefächertes Themenspektrum ab. Der nachfolgende Bericht fasst die Vorträge aus der Arbeitsgruppe *Modellierung nichtstandardisierter Schriftlichkeit* (AG 10) zusammen, zu der **Michael Beißwenger** (Dortmund), **Stefanie Dipper** (Bochum), **Stefan Evert** (Erlangen) und **Bianka Trevisan** (Aachen) eingeladen hatten. Ziel dieser AG war es, Herausforderungen und Möglichkeiten bei der Erfassung und Analyse nichtstandardkonformer sprachlicher Strukturen in unterschiedlichen linguistischen und computerlinguistischen Forschungsfeldern darzulegen und die aus den verschiedenen Perspektiven stammenden Erkenntnisse, Verfahren und Lösungsansätze bereichsübergreifend zu diskutieren.

Die einzelnen Beiträge unterscheiden sich durch das untersuchte Medium, das Genre oder die Varietät. Mit Blick auf die primär behandelten Fragestellungen lassen sie sich grob in zwei Gruppen einteilen: die Gruppe theoretisch-linguistischer Arbeiten und die der methodisch-technischen. In der ersten Gruppe wurden varietäten- bzw. domänenspezifische Aspekte behandelt, die die Bereiche Graphematik und Orthographie, konzeptionelle Mündlichkeit, Dialekt- und Soziolektforschung, sowie Spracherwerbs- und Übersetzungsforschung betreffen. Die zweite Gruppe

umfasste Beiträge zur Evaluierung und Verbesserung automatischer Verfahren bei der computergestützten Analyse der genannten domänenspezifischen Texte.

Beiträge zum Thema Graphematik und Orthographie

Felix Bildhauer (Berlin) beschäftigte sich mit dem Thema „Majuskelschreibung in Webkorpora: Verteilung und Funktion“ in einem sehr großen (> 1 Mrd. Token), die Sprachen Deutsch, Schwedisch, Französisch und Spanisch umfassenden Webkorpus.[1] Für die computergestützte Aufbereitung von Webkorpora stellen Formen nichtstandardkonformer Orthographie auf der einen Seite zwar eine technische Herausforderung dar, da für die Verarbeitung des Korpus mit der gängigen NLP-Software sowie für eine bessere Durchsuchbarkeit des Korpus eine automatische Orthographienormalisierung vorgenommen und deren Zuverlässigkeit gewährleistet sein muss. Auf der anderen Seite sind jedoch nichtstandardkonforme Schreibweisen in Webkorpora höchst interessante Daten für linguistische Studien bspw. auf dem Gebiet der Informationsstruktur. Ziel ist es daher, die Originalschreibung neben der normalisierten Version beizubehalten, um so nicht nur deren Funktion und Verteilung analysieren zu können, sondern auch in längerfristiger Hinsicht ein informationsstrukturell annotiertes Korpus zu generieren. In diesem Sinn wurde anhand plastischer Beispiele gezeigt, wie wertvoll Majuskelschreibungen für die Analyse informationsstruktureller Kategorien wie Fokus sind.

Ulrike Sayatz und **Roland Schäfer** (Berlin) verwendeten in ihrer Analyse von „Klitika und Apostrophschreibung in Webkorpora zwischen Graphematik und Registerklassifikation“ ganz ähnliche Daten (DECOW2012). Anhand der Kurzformen des Indefinitartikels (*einen* vs. *nen* vs. *n*) wurde ein Phänomen konzeptioneller Mündlichkeit in Webdaten untersucht. Neben einer empirischen Analyse des Gebrauchs von Kurzformen im grammatischen und graphematischen System gingen die Vortragenden auch der Frage nach, inwiefern dieses Merkmal typisch für Forendiskussionen ist und wie gut solche Erkenntnisse zur automatischen Anreicherung von Korpora mit Metadaten und letztendlich auch zur Registerklassifikation beitragen können.

Mit noch nicht standardisierter Schriftlichkeit um 1650 beschäftigte sich **Sandra Waldenberger** (Bochum) in ihrem Vortrag „Schriftsprachliche Variation und emergenter Standard im Übergang zur nhd. Orthographie“ und eröffnete somit eine sprachhistorische Perspektive. Neben einem Einblick in ihre Daten-

1 http://hpsg.fu-berlin.de/~rsling/downloads/pubs/SchaeferBildhauer_LREC2012_BuildingLargeCorpora.pdf

quelle, bestehend aus von unterschiedlichen Schreibern angefertigten, aber inhaltlich identischen, handschriftlichen Protokollen aus den „Akten des Westfälischen Friedens“, illustrierte sie die für die Untersuchung schriftsprachlicher Variation und die Beantwortung der Frage nach der Existenz eines „Proto-Standards“ notwendigen Prozesse der Datenaufbereitung von der Transkription der Handschriften über die Tokenisierung und Alignierung bis hin zur Lemmatisierung und somit Abstraktion von graphischer Variation. Die Präsentation des Forschungsziels und der technischen Herangehensweise wurde ergänzt durch einen Ausblick auf die Auswertungsperspektive. Anhand eines Beispielsatzes und der Analyse von Variationsfeldern, bspw. bei der Doppelkonsonantenschreibung, der Markierung von Vokallänge oder der Getrennt-/Zusammen- bzw. Groß-/Kleinschreibung, vermittelte Sandra Waldenberger einen Eindruck vom Ausmaß der graphematischen Variation an der Schwelle zum Neuhochdeutschen.

Mit „Andersschreibungen luxemburgischer Jugendlicher auf digitalen Pinnwänden“ beschäftigte sich **Luc Belling** (Luxemburg) und stellte damit Normabweichungen in Texten von Schülerinnen und Schülern in einer multilingualen und multikulturellen Situation vor. Eine zentrale Fragestellung des Vortrags war, ob die Normabweichungen der untersuchten Pinnwandeinträge die tatsächlichen Orthographiekenntnisse der Schüler widerspiegeln oder ob es sich bei den Abweichungen um eine bewusst eingesetzte Sprachvariation handelt. Die in der Untersuchung festgestellte hohe Fehlerquote wurde einerseits auf kreativen Sprachgebrauch andererseits aber auch auf Defizite im Bereich Orthographie zurückgeführt.

Meikal Mumin (Köln und Napoli „L’Orientale“; Titel: „Explaining the Unexplainable – On the Challenges of Transliterating Arabic Based Script“) beschrieb die Schwierigkeiten beim Transkribieren von historischen arabischen Schriftsystemen, die für die Verschriftlichung unterschiedlicher afrikanischer Sprachen verwendet werden. Sein Augenmerk richtete sich somit nicht auf eine linguistische Fragestellung zu den Korpusdaten, sondern auf die grundlegende Frage nach der Digitalisierung und korpuslinguistischen Weiterverarbeitung nicht-normierter arabischer Schriften. Verglichen mit deutschen Sprachdaten ähneln die beschriebenen Probleme am meisten denen der historischen Korpuslinguistik, gehen jedoch weit über die dortigen Tokenisierungs-, Transliterations- und Normalisierungsprobleme hinaus, da einige der verwendeten Zeichen nicht in Unicode repräsentiert sind, was daran liegt, dass einige Grapheme nicht nur Laute, sondern auch größere linguistische Einheiten wie z. B. ganze Phrasen repräsentieren können und die Linearität der Texte nicht immer stringent von rechts nach links gerichtet ist. Als eine Lösung wurde die Nutzung von Mehrebenenrepräsentationen der zu transkribierenden Texte vorgestellt. Ein solcher Ansatz ist unumgänglich, wenn nicht gänzlich von dem Wesen der einzelnen Schriften abstrahiert werden soll. Um das Schriftbild und individuelle Grapheme

gemeinsam mit den dadurch ausgedrückten lautlichen Sequenzen darstellen zu können, ist daher eine Alignierung unterschiedlicher (graphischer und phonetischer) Repräsentationen der Texte erforderlich.

Ebenfalls mit nichtstandardkonformer Orthographie beschäftigten sich **Christa Dürscheid** und **Simone Ueberwasser** (Zürich; Titel: „(Kein) liberaler Umgang mit der orthographischen Norm. Empirische Befunde zur schriftlichen Alltagskommunikation"), jedoch aus einer den vorausgegangenen Vorträgen entgegengesetzten Perspektive. In ihrer Studie anhand des Schweizer SMS-Korpus[2] ging es nicht darum, welche Abweichungen von der Standardorthographie vorzufinden sind, sondern explizit darum, welche Bereiche bei derartigen Abweichungen systematisch ausgeklammert werden und somit weitestgehend normstabil sind. Hierbei zeigten die Autorinnen unter anderem am Beispiel von *vielleicht*, dass die regelkonforme Orthographie beibehalten wird, wenn Abweichungen Regelunkenntnis vermuten lassen könnten.

Der Vortrag von **Klaus Geyer** (Odense) behandelte den Themenbereich „Dialektgraphien in Mundart-Ausgaben von Asterix-Comics". Anhand ausgewählter Alben zeigte er, welche Besonderheiten bei der Verschriftlichung von regionalen, zum Großteil noch nicht verschrifteten Sprachformen zum Ziel der heiteren Unterhaltung anzutreffen sind und wie sich die Übersetzungsstrategien unterscheiden können.

Beiträge zum Thema medial/konzeptionell mündlicher Sprachgebrauch

Christian Mair (Freiburg; Titel: „From transcription to evocation: digital/visual ethnolinguistic repertoires in Caribbean and West African CMC diasporas") stellte als eingeladener Gastredner Arbeitsmethoden und Ergebnisse bei der Erforschung webbasierter „Communities of Practice" im Bereich der karibischen und westafrikanischen Diaspora vor. Die Datenbasis bildeten drei große Korpora, bestehend aus Forenkommunikation von Online-Gemeinschaften, anhand derer sich nicht nur linguistische Strukturen untersuchen lassen, sondern die dank zahlreicher metasprachlicher Kommentare auch Aufschluss über implizit oder explizit zum Ausdruck gebrachte Sprachideologien liefern. Neben den Erwägungen bei der Korpuserstellung stellte Christian Mair zudem einige Visualisierungsmöglichkeiten des in Freiburg entwickelten Korpusprogramms N-CAT vor, mit dem sich auf einer Weltkarte die regionale Verteilung und Beteiligungsintensität sowie das Geschlecht der

2 http://www.linguistik-online.de/48_11/

Forenteilnehmer veranschaulichen lassen, um den virtuellen Raum im realen Raum zu verorten. Am Beispiel des „Corpus of Cyber-Jamaican“ (CCJ) zeigte er im Anschluss die verschiedenen Stufen auf, die die jamaikanische Kreolsprache beim Übergang von einer mündlichen zu einer visuellen Vernakulärsprache durchlaufen kann, d. h. welche Strategien die Teilnehmer einsetzen, um dem Bedürfnis nachzugehen, den Besonderheiten der von ihnen gesprochenen Sprache – und dadurch ihrer Identität – auch in ihrer schriftlichen, computervermittelten Form Ausdruck zu verleihen. Er illustrierte anhand zahlreicher Textbeispiele, dass dies nicht nur durch Transkription, d. h. die schriftliche Wiedergabe gesprochener Charakteristika, vollzogen werden kann, sondern auch durch den Einsatz eines bewusst informellen bzw. anti-formellen Sprachstils, also des Basilekts, insbesondere auch von Sprechern, die in normalen Sprechsituationen nur den Mesolekt gebrauchen würden. Weitere Strategien sind die Verwendung von „visual styling“, also einer Art Verfremdung des Textes durch idiosynkratische Schreibformen, oder aber eine Mischform der verschiedenen Strategien. Interessant ist hierbei, dass scheinbar nicht immer die Absicht verfolgt wird, auch tatsächlich die Aussprache wiederzugeben, sondern dass es sich bei den nichtstandardkonformen Schreibweisen vielmehr auch häufig um „eye dialect“ handelt, dessen primäre Funktion es ist, durch seine Abweichung von der Norm herauszustechen, und der es somit den Mitgliedern der Online-Gemeinschaft erlaubt, sich durch ihren Sprachgebrauch abzugrenzen. Anhand der vorgestellten Beispiele und seiner Typologie von Nichtstandard-Phänomenen legte Christian Mair nicht nur die stetig zunehmende Bedeutung und Diversifizierung von geschriebener Vernakulärsprache dar, sondern demonstrierte zudem, wie Sprache als eine Ressource genutzt wird, deren Charakteristika insbesondere in nicht reglementierten, multilingualen und multikulturellen Kontexten von den Sprechern bzw. Schreibern interaktiv ausgehandelt werden. Durch die systematische Korpusanalyse schriftlichen Sprachgebrauchs in webbasierten „Communities of Practice“ eröffnen sich somit vielversprechende Forschungsmöglichkeiten im Bereich der Sozio- und Varietätenlinguistik. Die sich anschließende Diskussion drehte sich in erster Linie um die besonderen Eigenschaften und den stetig wachsenden Stellenwert computervermittelter Kommunikation und die damit verbundenen Veränderungen im Sprachgebrauch, insbesondere was regionale Unterschiede und die Verschriftlichung von Vernakulärsprache anbelangt. An die Stelle regional bzw. territorial bestimmter, nach außen weitgehend abgegrenzter Sprechergemeinschaften sind inzwischen mobile Sprecher in stetig wechselnden bzw. sich verändernden Gemeinschaften getreten, die die Ressourcen ihrer Vernakulärsprache neu aushandeln und sie auch zunehmend in den mehr oder minder öffentlichen Schriftgebrauch übertragen.

Mit dem Beitrag „Äh… Ähm… Filled Pauses in Computer-Mediated Communication“ stellte **Ines Rehbein** (Potsdam) gemeinsame Arbeiten mit **Sören Scha-**

lowski, **Nadja Reinhold** und **Emiel Visser** (Potsdam) vor. Eine Analyse des Vorkommens dieser für gesprochene Sprache so typischen Interjektionen in Twitter-Nachrichten sollte Aufschluss darüber geben, ob Tweets trotz ihrer medial schriftlichen Realisierung konzeptionell mündlich ausgerichtet sind und inwiefern Gemeinsamkeiten bzw. Unterschiede hinsichtlich der Verteilung und Funktion gefüllter Pausen im Vergleich zur gesprochenen Sprache bestehen. Als Korpora für den direkten quantitativen und qualitativen Vergleich mit spontaner gesprochener Sprache dienten das Potsdam Twitter Corpus (PoTwiC) und das Kiezdeutsch-Korpus.[3] Ines Rehbein zeigte bei der Präsentation ihrer quantitativen Ergebnisse, dass sich die verschiedenen Funktionen von ihrer (relativen) Häufigkeit zueinander gleich verhalten (in absteigender Reihenfolge: Markierung von Häsitationen, Korrekturen, Abbrüche und Wiederholungen). Insgesamt betrachtet sind die Phänomene jedoch in den Twitter-Daten seltener. Vor allem sind Abbrüche und Wiederholungen hier selten anzutreffen. Betrachtet man die Distribution, stehen *äh* und *ähm* – zumindest in ihrer Funktion als Häsitationsmarker – in Tweets meist am Anfang oder Ende der Äußerung und nicht wie im gesprochenen Korpus häufig satzintern. Unterstützt durch zahlreiche Beispiele zeigte die Vortragende in ihrer qualitativen Analyse zudem, dass *äh* und *ähm* als Markierung von Häsitationen und Korrekturen im Twitter-Korpus bewusst eingesetzt und mit zusätzlichen Funktionen überlagert werden. So können Schreiber damit bspw. negative Gefühle zum Ausdruck bringen oder zu erkennen geben, dass ihre Äußerung mit Humor zu verstehen ist. Dies führte Ines Rehbein zu der Schlussfolgerung, dass mit *äh* und *ähm* hier zwar Mündlichkeit simuliert wird, diese Füllwörter jedoch bewusst eingesetzt werden und somit Bedeutungsstatus erlangen.

Thomas Schmidt (IDS Mannheim) stellte in seinem Beitrag „Orthographische Normalisierung und PoS-Tagging von Transkriptionen gesprochener Sprache“ die Schwierigkeiten und die erarbeiteten Lösungsansätze im Umgang mit der korpuslinguistischen Bearbeitung mündlicher Gespräche im FOLK-Korpus[4] (Forschungs- und Lehrkorpus gesprochenes Deutsch) des Instituts für Deutsche Sprache vor. Für die Aufbereitung gesprochensprachlicher Daten existieren noch keine automatischen Werkzeuge wie diejenigen, die in den Beiträgen zur Aufbereitung von Webkorpora vorgestellt wurden. Somit werden in FOLK medial mündliche Daten manuell mittels literarischer Transkription verschriftlicht und in einem zweiten Schritt semi-automatisch normalisiert, wobei mittels einer Grund- bzw. Wortformliste (DeReWo) des Deutschen Referenzkorpus DeReKo[5] und auto-

3 http://www.dsdigital.de/ce/das-kiezdeutsch-korpus/detail.html
4 http://agd.ids-mannheim.de/folk.shtml
5 http://www1.ids-mannheim.de/kl/projekte/korpora/

matischer Vorverarbeitung in einem speziell entwickelten Editor („OrthoNormal“) den Nichtstandardformen normalisierte Formen zugewiesen werden.

Beiträge zum Thema Lernersprache

In ihrem Vortrag „Why learner texts are easy to tag“ untersuchten **Marc Reznicek** (Berlin) und **Heike Zinsmeister** (Stuttgart) die Performanz automatischer Tagger auf Textdaten fortgeschrittener Lernender des Deutschen als Fremdsprache. Die Daten stammen aus einem Lernerkorpus, welches im Rahmen des Netzwerks „Kobalt-DaF“ erhoben wurde, und zeichnen sich durch spracherwerbsspezifische Besonderheiten und Fehler aus, welche für PoS-Tagger, die auf Standardsprache trainiert wurden, problematisch sein können. Um zu ergründen, wie groß der Einfluss der Lernersprache auf solche Analyseprobleme ist, wurden unbearbeitete Lerneressays mit dem TreeTagger[6] und entsprechende Korrekturfassungen (sog. Zielhypothesen) wortartenannotiert und die Taggingergebnisse anschließend verglichen. So zeigte sich, dass die Genauigkeit des Taggers auf Lernerdaten signifikant niedriger ist als auf Muttersprachlerdaten und dass die Genauigkeit auf der Ebene der Zielhypothesen gegenüber den ursprünglichen Lerneräußerungen durch die Korrektur spracherwerbsspezifischer lexikalischer Fehler und Wortstellungsfehler signifikant steigt. Die jeweiligen Performanzunterschiede sind mit 97,2 % Genauigkeit auf den muttersprachlichen Texten und 95,3 % auf den unkorrigierten Lernertexten jedoch deutlich geringer als erwartet.

Aivars Glaznieks, **Egon Stemle**, **Andrea Abel** und **Verena Lyding** (Bozen) stellten in ihrem Vortrag „Herausforderungen bei der Erstellung eines L1-Lernerkorpus: Lösungsvorschläge aus dem Projekt ‚KoKo‘“ die wesentlichen Schritte bei der Erstellung eines Korpus aus Schüleressays vor, das zur Untersuchung der Auswirkung diatopischer und biographischer Merkmale auf die Sprachkompetenz der ProbandInnen dienen soll. Die Sprache des Korpus ist Deutsch, der Erhebungsraum Südtirol, also ein mehrsprachiges Gebiet. Bei der Korpuserstellung orientierten sich die Mitwirkenden an bereits existierenden Projekten und vermieden oder antizipierten somit gewisse Probleme bei der Datenakquise und -aufbereitung. Da im Bereich der Lernerkorpora in den vergangenen Jahren vor allem L2-Korpora erstellt wurden, ist der Einfluss bspw. des englischen ICLE-Korpus oder des deutschen FALKO-Korpus erkennbar und führt so zu einer besseren Vergleichbarkeit der verschiedenen Korpora (bspw. liegen den erhobenen Essaytexten ähnliche Fragestellungen zugrunde).

6 http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

Beiträge zu übersetzungswissenschaftlichen Aspekten nichtstandardisierter Schriftlichkeit

Stella Neumann, **Paula Niemietz** und **Tatiana Serbina** (Aachen) behandelten in ihrem Vortrag „Linguistic Annotation of Text Fragments in a Keystroke Logged Translation Corpus" das Thema nichtstandardisierter Schriftlichkeit anhand von Korpusdaten, die die Entstehung von Übersetzungstexten abbilden. Die Daten entstanden durch die Aufzeichnung von Tastenanschlägen bei dem Prozess der Übersetzung wissenschaftlicher Texte vom Englischen ins Deutsche durch professionelle Übersetzer einerseits und Fachwissenschaftler (Physiker) andererseits. Die Endprodukte der Übersetzungen sind weitestgehend standardsprachlich; sie enthalten jedoch Performanzfehler, die durch den nachvollziehbaren Textproduktionsprozess erklärbar werden. So ist der Kongruenzfehler *ein**e** dünnes Blatt Alufolie* auf den Entstehungsprozess zurückzuführen, in welchem die Form *eine dünne Alufolie* vorformuliert wird. Der Entstehungsprozess der finalen Texte wird durch die Aufzeichnung von Einfügungen, Löschungen und Bewegungen von Zeichen(ketten) abgebildet. Das Referenzsystem ist somit eine Zeitleiste und die Texte werden als Folgen von Zeichen repräsentiert. Die Endversionen der Texte werden dagegen als fortlaufende Token repräsentiert. Das Korpus zeichnet sich also durch ein heterogenes Tokenkonzept und besonders komplexe Daten aus, deren linguistischer Nutzen vielseitig ist.

Beiträge zum Thema Analyse und automatische Aufbereitung von Webkorpora und von Sprachdaten aus Genres internetbasierter Kommunikation

Die eingeladene Sprecherin **Jennifer Foster** (Dublin) stellte in ihrem Vortrag „#hardtoparse: The Challenges of Parsing the Language of Social Media" drei in den Jahren 2010/11 erarbeitete Studien zur Evaluation und Verbesserung von Parsern für die syntaktische Analyse von Texten aus dem Bereich der internetvermittelten Kommunikation vor. Allgemein gilt für das Englische, dass anhand von Zeitungstexten trainierte Parser bei Texten internetvermittelter Kommunikation rund 10 % schlechtere Ergebnisse erzielen. Um diesen Einbußen entgegenzuwirken, können unter anderem die Trainingsdaten oder die zu analysierenden Daten angepasst werden. Die Anpassung des Parsers wird als Domänenadaptation bezeichnet, die Anpassung der Analysedaten als Normalisierung. Zunächst zeigte Foster, wie eine Domänenanpassung bei der Analyse von Forentexten erfolgen kann und welchen Effekt sie hinsichtlich der Analyseergebnisse hat. Hierfür wurde zunächst untersucht, welche Strukturen in Forentexten die Analyse des Parsers

besonders negativ beeinflussen, um daraus Anpassungen der Trainingsdaten und der Forentexte abzuleiten. Die Forendaten wurden in erster Linie auf drei Merkmale hin bearbeitet: Großschreibungen und Binnenmajuskeln wurden aufgehoben, forentypische Abkürzungen und Kurzformen (wie *cos* für *because*) wurden aufgelöst und diskursmarkierende Akronyme wie *lol* gelöscht. Dies erfolgte großteils automatisiert. Zudem wurde das in den Daten häufig fehlende Adverbsuffix *-ly* an adverbial gebrauchten Adjektiven entfernt. Die Verbesserung durch diese Maßnahmen ist signifikant, gleicht jedoch nur etwa ein Viertel der eingangs beschriebenen Einbußen aus. Im weiteren Verlauf des Vortrags folgte eine differenziertere Darstellung des beschriebenen Ansatzes zur Verbesserung der Parserleistung, wobei unterschiedliche Parser und unterschiedliche Anpassungs- und Trainingsverfahren miteinander kombiniert und verglichen sowie unterschiedliche Webgenres differenzierter betrachtet wurden. So wurden bspw. die zwei Trainingsverfahren „self-training" und „up-training" anhand von Wall-Street-Journal- und Twitter-Daten verglichen und evaluiert. Bei dem ersten Verfahren werden automatische syntaktische Analysen als zusätzliche Trainingsdaten für denselben Parser verwendet, der sie erzeugt hat; beim „up-training" werden die Analysen eines anderen Parsers als Trainingsdaten verwendet. Das „self-training" lieferte jeweils bessere Ergebnisse für beide Datentypen, wobei der oben genannte Unterschied von 10 % zwischen Zeitungsdaten und domainfremden Daten etwas verkleinert wurde (die F-Score-Ergebnisse liegen bei 83,8 % für Daten des Wall Street Journal und bei 76,3 % für die Twitter-Daten). Der anschließende Vortragsteil widmete sich weiteren, differenzierteren Webgenres (Blogs, Newsgroups, Reviews, Answers und Emails) sowie einem Vergleich von Konstituenten- mit Dependenzparsern. Hierbei stellte Jennifer Foster ein komplexes Vorgehen dar, das den in den Experimenten verwendeten domänenspezifischen Trainingsprozess sowie den Test-Prozess beschreibt. Das Ergebnis der Versuchsreihe ist eine differenzierte Übersicht, in der die Testergebnisse von zwei Konstituenten- und einem Dependenzparsingverfahren gegenübergestellt werden, indem die Ergebnisse dieser drei Verfahren bezüglich der fünf genannten Webgenres verglichen und zusätzlich mit Parsingergebnissen für das Wall Street Journal kontrastiert werden. Es ergibt sich für die Reihenfolge der Webgenres nach ihrer Analysequalität das folgende Bild: Wall Street Journal > Blogs > [Newsgroups, Reviews] > [Answers, Emails]. Bei diesem Ranking gilt der 10 %-Unterschied zwischen Zeitungsdaten und Webdaten nur noch für den Vergleich von Wall Street Journal und den Webgenres mit den schlechtesten Parsingergebnissen (Answers und Emails).

In ihrem Vortrag „Korpusbasierte Analyse internetbasierter Kommunikation" stellten **Thomas Bartz** und **Angelika Storrer** (Dortmund) die Herausforderungen bei der Erstellung des „Deutschen Referenzkorpus zur internetbasierten Kom-

munikation“ (DeRiK)[7] dar. Am Beispiel von Daten aus Wikipedia-Diskussionen und dem Dortmunder Chatkorpus wurde ermittelt und evaluiert, welche Probleme entstehen, wenn die Daten mit automatischen Verfahren analysiert werden, die an standardsprachlichem Material entwickelt wurden. Konkret ging es um Ergebnisse, die aus automatischem Tokenisieren und Wortartentagging mittels TreeTagger sowie OpenNLP-Tagger[8] resultieren. Problematisch sind hierbei u. a. die irreguläre Verwendungen von Spatien sowie zusammengezogene Formen wie *haste* oder *meinste*. Davon abgesehen gibt es Formen, die zwar korrekt tokenisiert sind, die jedoch vom Tagger nicht den dafür im Tagset vorgesehenen Klassen zugewiesen werden. Noch problematischer sind jedoch Formen, für die in den vorhandenen Tagsets keine adäquaten Kategorien existieren, was z. B. für Emoticons und Aktionswörter (*grins*, *freu*, *nachdenklichguck*) gilt. Fazit ist, dass existierende Tagsets in Bezug auf Phänomene internetbasierter Kommunikation erweitert, computerlinguistische Verfahren für die Tokenisierung und das Wortartentagging an die Besonderheiten internetbasierter Schriftlichkeit angepasst und geeignete Standards für die Annotation solcher Daten etabliert werden müssen.

In ihrem Beitrag „Linguistic Annotation of Computer-Mediated Communication – (not only) an Explorative Analysis“ behandelten **Kay-Michael Würzner** (Potsdam), **Lothar Lemnitzer**, **Alexander Geyken** und **Bryan Jurish** (Berlin) ebenfalls die Problematik der Aufbereitung von internetbasierten Korpusdaten. Wie im Vortrag von Thomas Bartz und Angelika Storrer ging es um das Vorhaben, Daten aus dem DeRiK-Projekt mit automatischen Verfahren zu tokenisieren, zu lemmatisieren und mit Wortarten zu annotieren. Der Beitrag konzentrierte sich auf die Evaluation unterschiedlicher statistischer Methoden zur Tokenisierung auf Wort- und Satzebene und der Zuweisung von Wortartentags. Für die Tokenisierung und Satzsegmentierung wurde ein auf Hidden-Markov-Modellen basierender Tagger verwendet („moot“) und dessen Leistung anhand eines manuell tokenisierten Subsets (100.000 Token) des Dortmunder Chatkorpus evaluiert. Während die Tokenisierung mit zuverlässiger Präzision funktionierte (über 97 %), wurden nur ca. 65 % der nicht markierten Satzendungen erkannt. Für die Wortartenannotation wurden der „moot“-Tagger und der TreeTagger miteinander verglichen. Die Ergebnisse sind relativ ähnlich, blieben jedoch mit ca. 75 % Korrektheit weit hinter den Leistungen aktueller Tagger auf standardsprachlichen Texten zurück.

Markus Dickinson, **Mohammad Khan** und **Sandra Kübler** (Indiana) behandelten in ihrem Beitrag „Towards Parsing YouTube Comments“ wie schon Jennifer

7 http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/derik-a-german-reference-corpus-of-computer-mediated-communication/

8 http://opennlp.apache.org/documentation.html

Foster die syntaktische Analysierbarkeit von Textdaten aus dem Bereich der internetbasierten Kommunikation. Die Untersuchungen konzentrierten sich auf die Eigenschaften von YouTube-Kommentaren, die die Performanz der auf Standarddaten trainierten Parser verschlechtern, wie Emoticons, internetspezifische Abkürzungen, anderssprachiges Material, Abweichungen von der orthographischen Norm und Kontraktionen. Für jede dieser Kategorien wurde eine automatisierte Strategie entwickelt, um die YouTube-Kommentardaten mit herkömmlichen Parsern (die auf Zeitungstexten aus der Penn Treebank trainiert wurden) besser analysieren zu können. Die Schlussfolgerung aus diesen Studien war, dass häufig relativ einfache Methoden ohne linguistische Regeln den besten Effekt lieferten. So wurde bspw. nicht-englischsprachiges Material in den untersuchten YouTube-Kommentaren am wirkungsvollsten eliminiert, wenn nicht-alphabetische Zeichen aus den Daten entfernt wurden (die meisten anderssprachigen Kommentare in den behandelten Daten enthalten asiatische Schriftzeichen).

Sämtliche Vorträge aus der AG wurden im Anschluss an die Veranstaltung online dokumentiert: http://www.sfb632.uni-potsdam.de/dgfs-2013/AG_10.html