

Anmerkungen zur scheinbaren Konkurrenz von numerischen und symbolischen Verfahren in der Computerlinguistik¹

1 Einleitung

Veränderungen in der Mode vollziehen sich schrittweise, allmählich, oftmals unmerklich. Manchmal kommt es dann doch zu einem plötzlichen Bruch, zu einer eindeutig bestimmten Opposition zwischen dem modischen Jetzt und seinem Vorgänger. Deutlich erkennbar etwa, als um die Wende von den 70er zu den 80er Jahren die weite Karotte die enge Schlaghose verdrängte und im Gefolge dieses Wechsels taillierte Sakkos und Hemden *big suits* und Polohemden weichen mussten.² Etwas Ähnliches hat sich vor einigen Jahren in der Computerlinguistik ereignet: Wurde diese seit Beginn der 80er Jahre durch deduktive, regelbasierte Verfahren beherrscht, so gab es seit Beginn der 90er Jahre zunächst eine Hinwendung, schließlich Mitte der 90er Jahre eine Flucht zu induktiven, numerisch basierten Verfahren, die mittlerweile die Computerlinguistik nahezu vollständig zu beherrschen scheinen.³

Eingeleitet wurde dieser Wechsel durch ein von bestimmten Enttäuschungen in der übermäßigen Verwendung logischer, aber ineffizienter Formalismen ausgelöstes, erleichtertes ‚*Es geht auch anders.*‘; dies aber schlug recht zügig in ein ‚*Es geht nicht anders.*‘ um. Dieser methodischen Eindeutigkeit folgend ignoriert man die Evaluationskriterien für regelbasierte Systeme und richtet sich statt dessen an den Evaluationsmethoden statistischer Verfahren aus, die im allgemeinen eine Minimalanforderung, eine *Baseline*, vorgeben und die Systemperformance danach bemessen, wie weit sie sich von dieser Minimalanforderung weg und zu 100 % hin bewegt. Andere Kriterien werden hingegen ignoriert, so beispielsweise die Plausibilität der Modellierung. So sind N-gramm-Modelle erkennbar keine korrekten Modelle für sprachliche Abhängigkeiten, weil Sprache kein stationärer Prozess ist.⁴ Wäre dies der Fall, so gäbe es weder die Auslautverhärtung, noch die in den Sprachen der Welt mit gewisser Regelmäßigkeit beobachtbare Positionierung finiter Verben an *bestimmten* Positionen. Aber solcherlei theoretischer Ballast verschreckt niemanden, solange die Verfahren funktionieren.

Dies hat zur Folge, dass man gemeinhin gar nicht mehr fragt, ob die Verfahren tatsächlich funktionieren, ob also die so erzielten Ergebnisse der statistischen Verfahren ihren Opponenten tatsächlich überlegen sind. An die Stelle dieser Überlegung tritt, wenn überhaupt, eine Verbeugung vor dem neuen Paradigma, verbunden mit dem Wunsch, ob denn nicht die Unterlegenen doch auch einmal wieder mitspielen dürfen, d.h. ob nicht eine Integration des gar nicht so alten *ancien régime* durchaus sinnvoll wäre.

Diesen Fehler möchte ich nicht wiederholen, sondern stattdessen einen Schritt zurückgehen und fragen, ob denn tatsächlich die statistischen Verfahren *besser* sind als die symbolischen. Ich möchte dies an einer Anwendung verdeutlichen, die in jeder Hinsicht typisch für diese Fragestellung ist, der kategorialen Annotation von Wörtern in einem Korpus, dem sog. *Part-of-Speech-Tagging* (*POS-Tagging*, oder kurz *Tagging*). Beim Tagging haben wir es mit einem klassischen Ambiguitätsproblem zu tun. In Isolation betrachtet, kann nicht jeder Worttyp

¹ Ich danke Jan Strunk und Edgar Rudolph für ihre Mithilfe.

² Das Aufeinanderprallen dieser Richtungen kann man sich vergegenwärtigen, wenn man eine frühe Folge der Serie *Dallas* und direkt anschließend den letzten Auftritt David Byrnes im Film *Stop Making Sense* betrachtet.

³ Die Computerlinguistik der 80er Jahre unterscheidet sich auch in ihrer Beziehung zur theoretischen Linguistik wesentlich von der heutigen Computerlinguistik. Zu Beginn der 80er Jahre entstand eine bis dato einmalige Übereinstimmung zwischen theoretischer Linguistik und Computerlinguistik, die in der theoretischen Linguistik – bis hin zu bestimmten Varianten der Prinzipien- und Parametertheorie – insbesondere durch die Hinwendung zum Repräsentationalismus sichtbar wurde. Mittlerweile hat sich die theoretische Linguistik deutlich von der Computerlinguistik wegbewegt, sodass der symbolischen Computerlinguistik heute ein wichtiger Verbündeter fehlt.

⁴ Ein Prozess ist dann stationär, wenn der Wert einer Zufallsvariablen von einem Vorgängerzustand abhängig ist, *unabhängig davon, wo sich die Zufallsvariable und der Vorgängerzustand in einer Kette von Zuständen befinden*. Der Wert ist also nur vom Vorgängerzustand und nicht von einem weiteren Kontext, z.B. der Silbengrenze abhängig. Dass es dann an der Coda und nicht am Onset zur Auslautverhärtung kommt, ist dann ebenso überraschend wie die Tatsache, dass Deutsch wie viele andere Sprachen eine Verbzweit-, nicht jedoch eine Verbdritt-, Verbviert- oder Verbfünftsprache ist. Fairerweise sollte nicht unerwähnt bleiben, dass das Standardwerk zu statistischen Verfahren in der Computerlinguistik, Manning/Schütze (1999), die inhärenten Beschränkungen solcher Modelle nicht unerwähnt lässt.

eindeutig kategorial bestimmt werden, wohl aber in den meisten Fällen in einem gegebenen Kontext. Tagging ist aber über diese linguistische Problemstellung wissenschaftssoziologisch einschlägig, denn interessanterweise wird Tagging als eine der Musteranwendungen für statistische Verfahren beschrieben, obwohl nach den statistischen Evaluationskriterien ein regelbasierter Tagger deutlich besser abschneidet als jeder numerisch operierende Konkurrent (vgl. dazu Manning/Schütze 1999:373 und Samuelsson/Voutilainen (1997)).

Um nun zu ermitteln, was es bedeutet, besser oder schlechter zu sein, halte ich es für sinnvoll, zunächst einmal die Kritikpunkte an den symbolbasierten Systemen zu betrachten und zu überprüfen, ob sie tatsächlich keine Anwendung auf die statistischen Verfahren finden. Ich denke, dass die folgenden drei Kritikpunkte zentral sind:

Mangelhafte Abdeckung: Diese liegt dann vor, wenn eine Komponente nur bestimmte Konstruktionen, nicht jedoch beliebiges sprachliches Material, so wie es etwa in einem Korpus vorliegt, verarbeiten kann. Die Konsequenz daraus ist, dass Systeme nicht robust auf ‚überraschende‘ Eingaben reagieren können. Statistische Verfahren hingegen werden als robust bezeichnet, denn die Systeme lassen sich nicht leicht überraschen und besitzen entsprechend eine sehr breite Abdeckung.

Aufwändige Erstellung: Die Erstellung regelbasierter Systeme ist aufwändig, weil das deduktive sprachliche Wissen explizit kodiert werden muss. Darüber hinaus neigen Autoren solcher Regelsysteme zu Idiosynkrasien, sodass ein Regelsystem ohne den dazugehörigen Autor nicht adaptiert werden kann.

Abhängigkeit von einzelsprachlichen Gegebenheiten: Deduktive Regelsysteme sind schließlich bezogen auf die Abdeckung einer einzelnen Sprache und somit inhärent einzelsprachabhängig.

Gegner deduktiver Systeme behaupten, dass – zumindest bestimmte – statistische Verfahren eine vollständigere Abdeckung besitzen, die aufwändige Erstellung von Regelsystemen nicht verlangen und schließlich ohne großes Aufheben auf andere Sprachen adaptiert werden können.

Eine kritische Betrachtung der neuen Verfahren zeigt allerdings auf, dass diese mit ganz ähnlichen Problemen zu kämpfen haben wie ihre klassischen Gegner und dass diese Entgegnung somit einer faktischen Grundlage entbehrt. Die natürliche Schlussfolgerung kann nicht sein, beide Ansätze zu verwerfen. Sinnvoll erscheint es vielmehr, diese scheinbar polaren Herangehensweisen im Bewusstsein ihrer Schwächen zu kombinieren.

2 Das Ambiguitätsproblem beim Tagging

Das Ambiguitätsproblem beim Tagging kann anhand einiger empirischer Untersuchungen verdeutlicht werden. Ich beziehe mich hierbei auf das *Wall Street Journal*-Korpus, das Bestandteil des ACL/DCI-Korpus ist und durch das *Linguistic Data Consortium* vertrieben wird. Dieses Korpus hat den Vorteil, vollständig kategorial annotiert zu sein, sodass es als Basis für eine Auswertung dienen kann. Darüber hinaus ist das Verfahren zur Annotation detailliert beschrieben (Marcus et al. 1993) und erlaubt so klare Rückschlüsse auf verbleibende Fehler und Probleme. Bei einem willkürlich ausgewählten Teil dieses Korpus mit 2.000.000 Wörtern ergibt sich, dass von den vorkommenden Wort-Typen nur etwa 11 % kategorial mehrdeutig sind, dass diese 11 % jedoch 41 % der vorkommenden Wörter in diesem Korpus ausmachen.

Ein Beispiel für eine solche Ambiguität bietet das deutsche Wort *dem*, wie in (1) dargestellt.⁵

- (1) Was sagst Du zu *dem*(PDS), was Deine Mutter *dem*(ART) Fleischer gesagt hat, *dem*(PRELS) sie zufällig begegnet ist?

In Isolation betrachtet kann nicht entschieden werden, ob *dem* als Demonstrativpronomen (PDS), Relativpronomen (PRELS) oder Artikel (ART) klassifiziert werden sollte, in den o.g. Kontexten ist die Desambiguierung jedoch jeweils eindeutig.

Ein Vergleich mit anderen Korpora zeigt, dass diese Werte als repräsentativ gelten können. Ein Tagger muss somit bei ca. 40 % der Wortvorkommen in einem Korpus eine Ambiguitätsauflösung durchführen. Bei den folgenden Überlegungen gehe ich basierend auf dem o.g. Testkorpus von der folgenden Minimalanforderung aus: Wie hoch ist die Fehlerrate beim Tagging, wenn für jedes mehrdeutige Wortvorkommen im Korpus jeweils einfach das wahrscheinliche kategoriale Etikett gewählt wird? Das wahrscheinlichste kategoriale Etikett ist dasjenige, das im korrekten Fall am häufigsten zugewiesen wird. Es zeigt sich, dass man – bezogen auf die Annotation aller, d.h. mehrdeutigen und nicht mehrdeutigen, Wörter im Testkorpus – eine Fehlerrate von 4,9 % erhält, wenn man dieses Verfahren wählt. Betrachtet man nur die ambigen Wörter, so erhält man eine Fehlerrate von knapp 12 %. Da die Evaluationsergebnisse von Taggern jedoch zumeist auf die Annotation aller Wörter bezogen sind, setze ich die Fehlerrate von 4,9 % als Minimalanforderung an ein Taggingssystem an. Dies bedeutet umge-

⁵ Für die Annotation deutscher Korpora verwendete ich das STTS-Tagset, das als Standardtagset für die Annotation deutscher Korpora gilt (vgl. Schiller et al. 1999).

kehrt, dass ein Tagger wenigstens 95,1 % der Wörter korrekt annotieren muss, um dieser Minimalanforderung zu genügen.

3 Der Aufwand

Die in Abschnitt 2. vorgestellten Überlegungen vorausgesetzt, können wir nun unmittelbar das Problem des Aufwands diskutieren. Bei der Definition der Minimalanforderung bin ich natürlich davon ausgegangen, dass das Tagging bereits erfolgt ist. Normalerweise ist dies nicht der Fall, ein Tagger wird ja gerade zur Annotation verwendet. Aber auch in diesem Fall muss ein sog. Referenzkorpus vorliegen, das ein statistisch basierter Tagger zum *Training* benötigt. Es gibt unterschiedliche Möglichkeiten, ein Referenzkorpus zu erstellen. So kann man ein Referenzkorpus vollständig von Hand erstellen oder man verwendet einen Minimaltagger, mit dem eine initiale Annotation des Referenzkorpus vorgenommen wird, die dann wiederum manuell korrigiert wird.⁶ Für statistische Taggingverfahren ist die Güte des Referenzkorpus entscheidend, somit darf ein Referenzkorpus nicht zu viele Fehler enthalten. Ist die Fehlerrate zu hoch, so führt das Training zu Annahmen, die wahrscheinlich die Minimalanforderungen unterschreiten werden, wie Manning/Schütze (1999:205) bemerken: „*poor estimates of context are worse than none*“.

Die Erstellung eines Referenzkorpus setzt somit natürlich auch die Entwicklung eines entsprechenden kategorialen Inventars, des sog. *Tagsets* voraus. Dieses muss, damit es den Anforderungen der Abdeckung entspricht, die Sprache umfassend erfassen. Es ist notwendig, dass jedes Wort eines Korpus zumindest eine plausible Zuordnung erfährt. Daher sollte es nun ganz offensichtlich sein, dass die Erstellung des Referenzkorpus und die vorausgesetzte Erstellung des Tagsets nicht weniger aufwändig sein kann als die Erstellung entsprechender deduktiver Regeln. Der wesentliche Unterschied liegt weniger im Aufwand als vielmehr in der Kodierung: In einem regelbasierten System werden die Regeln explizit, *als Regeln* kodiert, in einem statistisch basierten System werden die Regeln implizit kodiert. Die angewandten Regeln müssen somit bei der Erstellung eines Referenzkorpus rekonstruiert werden, wie etwa die Handreichungen zeigen, die dem STTS-Tagset beigegeben werden (Schiller et al.1999). Hinzu kommt, dass man nicht davon ausgehen sollte, dass das Kategorisierungsproblem an sich gelöst ist. Als Beispiel hierfür geben Rudolph et al. (2001) die Klassifikation von Determinativkomposita an, deren Erstglied ein Eigenname ist. Schiller et al. (1999) schlagen vor, dass solche Komposita nach der Regel in (2) zu annotieren, also als Appellativa zu klassifizieren sind:

(2) NE (Eigenname) + NN („normales“ Nomen) → NN

Diese Regel macht allerdings bei Komposita wie *Gazastreifen* keinen Sinn, denn hierbei handelt es sich ja nicht um Streifen aus Gaza, sondern um *den* Gazastreifen, der folglich als Eigenname klassifiziert werden sollte. Gleiches gilt etwa für die *Ostsee*, das nach (2) ebenfalls als Appellativum annotiert werden sollte. Schließlich kann man die Regel in (2) nicht einfach dadurch modifizieren, dass man das fragliche Nomen nur dann als NE klassifiziert, wenn es sich um eine einmalige Entität handelt, denn in diesem Fall würde wohl auch die *Gretchenfrage* zum Eigennamen.

Aus dieser Perspektive ist für mich schwerlich erkennbar, dass der Aufwand für ein statistisch basiertes, flaches System tatsächlich geringer sein soll als der vergleichbare Aufwand für ein regelbasiertes System, dies um so mehr, als beide ja im Endeffekt auf der Grundlage von/nach bestimmten Regeln operieren.⁷

4 Abhängigkeit von einzelsprachlichen Gegebenheiten

Eine weitere Konsequenz der Abhängigkeit des Tagging von der Erstellung eines Referenzkorpus für das Training ist die inhärente Abhängigkeit von einer Einzelsprache. Will man einen gegebenen Tagger auf eine andere Sprache anwenden, so ist zunächst ein *neues* Tagset festzulegen, denn das kategoriale Inventar der einen Sprache

⁶ Wie Marcus et al. (1993) berichten, ist die Korrekturmethode der vollständigen Annotation vorzuziehen. Dies haben wir in verschiedenen Experimenten im Rahmen der Erstellung eines deutschen Referenzkorpus nachvollziehen können (vgl. Rudolph et al. 2001).

⁷ Diese Kritik gilt auch für Verfahren, die nicht auf einem Referenzkorpus aufsetzen, sondern eine Variante des Forward-Backward-Algorithmus für das Training verwenden. Auch hier ist zumindest eine lexikalische Zuordnung oder ein heuristisches Regelwerk erforderlich, um eine initiale Klassifikation durchzuführen. Vgl. Manning/Schütze (1999:357-361) und Brill (1995:558ff.).

stimmt nicht notwendig mit dem kategorialen Inventar einer anderen Sprache überein. Dies hat auch wesentlich damit zu tun, dass automatisches Tagging nur dann erfolgreich ist, wenn das kategoriale Inventar klein ist. Mehr als ca. 60-80 Tags sollten nicht verwendet werden (vgl. Marcus et al. 1993, Manning/Schütze 1999), weil ansonsten ein Datenproblem entsteht: Bei einer höheren Kardinalität der Kategorienmenge liegen immer weniger statistische signifikante Daten vor, ein dies nicht kompensierender, immer höherer Bedarf an Referenzkorpora entsteht.⁸ Dies bedeutet natürlich, dass eine wirkliche Ausdifferenzierung von Kategorien nicht vorgenommen werden kann und unterschiedliche Ausprägungen vielmehr zusammengelegt werden müssen. Dass eine solche Zusammenlegung etwa morphologischer Ausprägungen nicht so einfach auf eine andere Sprache übertragen werden kann, sollte unmittelbar einleuchten.

Darüber hinaus muss dann natürlich ein neues Referenzkorpus erstellt und neu trainiert werden. Eine Unabhängigkeit von einzelsprachlichen Ausprägungen ist nur dann gegeben, wenn das Verfahren tatsächlich *kein Training* benötigt.⁹ Dies ist jedoch beim Tagging nicht der Fall. Somit ist ein Tagger immer abhängig von einzelsprachlichen Besonderheiten und muss auf neue Sprachen auch neu adaptiert werden.

5 Abdeckung und Evaluation

Ein zentrales Argument gegen die Verwendung regelbasierter Systeme ist der durch Regeln zu erreichende Abdeckungsgrad. Man geht gemeinhin davon aus, dass durch ein statistisches Taggingverfahren eine korrekte Abdeckung in 95 bis 97 % der Fälle erreicht wird. In anderen Worten: In einem getaggtten Korpus ist mit einer Fehlerrate zwischen 3 und 5 % zu rechnen. Nun haben wir in Abschnitt 2. bereits gesehen, dass die Minimalanforderung für das von uns zur Evaluation verwendete Korpus bei 4,9 % lag. Wir haben, um dies zu überprüfen, zwei Tagger, den sog. Brill-Tagger (Brill 1995) und den Tagger TnT (Brandts 2000) mit unterschiedlich großen Trainingskorpora trainiert und anschließend anhand des bereits angesprochenen Testkorpus evaluiert. Die erzielten Ergebnisse sind, wie die Tabellen in (3) und (4) zeigen, zumindest beim Brill-Tagger problematisch.

- (3) Evaluation des TnT-Taggers unter Verwendung von Trainingskorpora einer Größe von 90.000 bis 450.000 Wörtern

Größe des Trainingskorpus	Prozentsatz korrekter Annotationen im Testkorpus
90.000	95,35 %
450.000	96,70 %

- (4) Evaluation des Brill-Taggers unter Verwendung von Trainingskorpora einer Größe von 90.000 bis 450.000 Wörtern

Größe des Trainingskorpus	Prozentsatz korrekter Annotationen im Testkorpus
90.000	93,13 %
450.000	94,33 %

Hierbei ist zu beobachten, dass der Brill-Tagger auch bei einem Referenzkorpus von 450.000 Wörtern eine Abdeckung zeigt, die die Minimalanforderung unterschreitet, während der TnT-Tagger die Minimalanforderung bereits bei 90.000 Wörtern übertrifft. Brill (1995) erzielt Ergebnisse von über 96 %, allerdings auf der Basis eines Trainingskorpus mit insgesamt 950.000 Wörtern – anbei dieselbe Quelle, die auch wir für unsere Evaluation verwendet haben, das *Wall Street Journal*-Korpus. Das relativ schlechte Abschneiden des Brill-Taggers kann auch nicht auf eine hohe Fehlerrate im Referenzkorpus selbst zurückgeführt werden, schließlich wurden beide

⁸ Bei einem Tagset von 50 Kategorien müssen bei einem Trigramm-Tagger bereits Vorhersagen über $50^3 = 125.000$ mögliche Übergänge getroffen werden; würde man die dreifach höhere Anzahl von Kategorien wählen, ergäbe sich das 27-fache an Übergängen, d.h. 3.375.000. Bei einem Trigramm-Modell errechnet sich die Steigerungsrate zwischen einem Tagset der Kardinalität N und einem der Kardinalität M durch $(M/N)^3$. Entsprechend verwendet das Penn-Tagset (Marcus et al. 1993) ebenso wie STTS (Schiller et al. 1999) zwischen 50 und 55 Tags.

⁹ Ein Beispiel für ein solches, trainingsfreies Verfahren präsentiert etwa der Ansatz zur Erkennung von Abkürzungen in Kiss/Strunk (2002a, b). Wenn für das Training nur sehr geringe unannotierte Datenmengen (< 20 kB) benötigt werden, wie etwa im Ansatz zur Spracherkennung bei Cavnar/Trenkle (1994), so kann man ebenfalls von einer Sprachenunabhängigkeit sprechen. Dieser Kritikpunkt richtet sich somit nicht gegen numerische Verfahren per se, sondern gegen die Annahme, dass diese prinzipiell sprachenunabhängig sind.

Tagger mit diesem Korpus gefüttert.¹⁰ Man kann also festhalten, dass die Verwendung des Brill-Taggers die vorherige, zu höchstens 3 % fehlerbehaftete Annotation eines mindestens 900.000 Wörter großen Korpus voraussetzt. Nun berichten Samuelson/Voutilainen (1997), dass ihr *regelbasiertes* System eine Fehlerrate von ca. 1 % besitzt, somit also die o.g. Ansätze um Längen schlägt.

Schließlich sei gefragt, was denn eine Fehlerrate von 5 % tatsächlich bedeutet. Dies lässt sich leicht anhand des Problems der Spracherkennung verdeutlichen. Eine Fehlerrate von 5 % bedeutet ja nichts anderes, als dass ein Geschäftsbericht mit 2.000 Wörter 100 Tippfehler enthält. Ein Sekretär mit diesen Fähigkeiten würde wahrscheinlich seine Arbeit nicht lange behalten und so kann es auch nicht verwundern, dass Programme zur Spracherkennung keine wirklichen Verkaufsschlager wurden. Bei einem annotierten Korpus mit 2.000.000 Wörtern fänden wir somit ungefähr 100.000 falsch annotierte Wörter. Hierbei stellt sich die Frage, ob diese in absoluten Zahlen doch recht beeindruckende Fehlerrate nicht vielleicht dadurch kompensiert werden könnte, dass die nachfolgenden Systeme die Fehler der vorhergehenden Systeme evtl. ‚schlucken‘. Leider kann ich hier zum Tagging keine Ergebnisse vorlegen, aber Kiss/Strunk (2002b) zeigen, welch starken Einfluss Fehler in der Satzgrenzenerkennung auf das anschließende Tagging haben: Durch eine präzisere Satzgrenzenerkennung kann die Fehlerrate des Taggers um ca. 75 % reduziert werden. Im Umkehrschluss bedeutet dies wohl doch, dass nachfolgende ‚robuste‘ Komponenten von qualitativ hochwertigem Input abhängig sind.

6 Schluss

Man kann wohl abschließend festhalten, dass von einer Überlegenheit statistischer Verfahren zumindest im Bereich des Tagging eigentlich nicht gesprochen werden sollte. Darüber hinaus muss die Opposition zwischen regelbasierten und numerischen Verfahren hier aufgeweicht werden, denn auch die statistischen Verfahren verwenden Regelsysteme. Selbst beim Lernen ohne Referenzkorpus ist ja zumindest eine Zuordnung der Wörter zu einem Lexikon bzw. auch eine heuristische Erkennung unbekannter Wörter nach Regeln notwendig. Statistische Verfahren haben – und dies wurde hier wahrscheinlich nicht ausreichend betont – durchaus ihre Berechtigung, sie sind nützlich; sie gestatten, insbesondere im Vergleich zur Introspektion, eine unmittelbarere und breitere Heranführung an das Phänomen Sprache. Die vorhandenen umfangreichen elektronischen Korpora verlangen nahezu danach, Sprache *auch* mit statistischen Mitteln zu untersuchen. Allerdings können die statistischen Verfahren die regelbasierten Verfahren nicht ersetzen. Somit muss dem Diktum vom *Es geht nicht anders*. deutlich widersprochen werden. Dass die statistischen Verfahren zur *Zeit so en vogue* sind und die regelbasierten Verfahren aussehen lassen wie eine alte Dallas-Folge, mag wohl auch daran liegen, dass zu viele Vertreter des alten Paradigmas nicht die Energie aufbringen, sich dem neuen Paradigma so weit zu öffnen, dass eine kritische Auseinandersetzung mit dem neuen auf der Basis des alten möglich wird. Die Mathematik ist eine geachtete, weil schwierige Wissenschaft, die statistische Sprachverarbeitung ist eine gefürchtete, weil in ihren Eigenschaften oftmals nicht gründlich genug betrachtete Disziplin.¹¹

¹⁰ Nach Marcus et al. (1993) liegt im Wall Street Journal-Korpus die Fehlerrate wahrscheinlich bei ca. 3 %.

¹¹ Den ersten Halbsatz dieses Satzes schulde ich der Neuen Zürcher Zeitung vom 10.10.2002, S. 34.

Literaturangaben:

- Brants, T. (2000): *TnT - A Statistical Part-of-Speech Tagger*. Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000, Seattle, WA.
- Brill, E. (1995): *Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging*. Computational Linguistics 21, S. 543-565.
- Cavnar, W.B./J.M. Trenkle (1994): *N-Gram-Based Text Categorization*. Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval
- Kiss, T./J. Strunk (2002a): *Scaled log likelihood ratios for the detection of abbreviations in text corpora*. In: Tseng, S.-C. (Hrsg.): Proceedings of COLING 2002. Taipei, S. 1228-1232.
- (2002b): *Viewing sentence boundary detection as collocation identification*. In: Busemann, S. (Hrsg.): Konvens 2002 Tagungsband. DFKI, Saarbrücken, S. 75-82.
- Manning, C.D./H. Schütze (1999): *Foundations of Statistical Natural Language Processing*. Cambridge/London: The MIT Press.
- Marcus, M./B. Santorini/M.A. Marcinkiewicz (1993): *Building a large annotated corpus of English: The Penn Treebank*. Computational Linguistics 19, S. 313-330.
- Rudolph, E./R. Pegam/S. Polubinski (2001): *Probleme bei der Erstellung eines deutschen Referenzkorpus für automatisches Tagging*. Manuskript, Sprachwissenschaftliches Institut der Ruhr-Universität Bochum.
- Samuelsson, C./A. Voutilainen (1997): *Comparing a linguistic and a stochastic tagger*. ACL 35/EACL 8, S. 246-253.
- Schiller, A./S. Teufel./ C. Stöckert/C. Thielen. (1999): *Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset)*. IMS, Universität Stuttgart, August 1999. <<http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-1999.pdf>>