Christian Mair

1 Erfolgsgeschichte Korpuslinguistik?

Überlegungen zum Fortschritt in der Sprachwissenschaft

Abstract: Der Beitrag gibt einen Überblick über die fünfzigjährige Erfolgsgeschichte der Korpuslinguistik. Er würdigt die enormen technischen Fortschritte und die quantitativ und qualitativ hervorragenden Korpus-Ressourcen, wie sie heute – zumindest für einige große Sprachen – zur Verfügung stehen. Eine direkte Folge der technischen Entwicklung ist der Aufstieg gebrauchsbasierter theoretischer Modelle in der wissenschaftlichen Sprachbeschreibung. Gleichzeitig weist der Beitrag auf die nach wie vor defizitäre Behandlung gesprochener Spontansprache in der Korpuslinguistik hin und empfiehlt, in Zukunft dem Aufbau mehrsprachiger Korpus-Ressourcen mehr Aufmerksamkeit zu schenken als bisher. Abschließend stellt er die Frage nach der zukünftigen Funktion der Korpuslinguistik im Rahmen der *Digital Humanities*.

Keywords: Digital Humanities, gesprochene Sprache, Korpuslinguistik, Mehrsprachigkeit

1 Einleitung

Korpuslinguistik und Konkordanzen haben eine lange Tradition, die weit in die Zeit vor der Digitalisierung zurückreicht. Ohne technische Unterstützung war diese Art von textwissenschaftlicher Analyse allerdings derart aufwendig, dass man die Mühen nur selten auf sich nahm. Konkordanzen wurden typischerweise dann erstellt, wenn – wie im Falle der Bibel oder der Werke Shakespeares – der ideelle Wert des Korpus den Aufwand rechtfertigte.[1] Forscher wie der amerikanische Strukturalist Charles Carpenter Fries, der sich mit derselben Sorgfalt und Andacht einem Korpus von transkribierten Gesprächen von Durchschnittsamerikanerinnen und -amerikanern aus dem Mittleren Westen

1 Als Pionier auf dem Gebiet der Bibelkonkordanz gilt der Dominikaner Hugo von St. Cher (gest. 1263) mit seinen „Concordantiae Sancti Jacobi". Die erste vollständige Konkordanz zu Shakespeares Werken ist Bartlett (1913).

Christian Mair, Englisches Seminar, Universität Freiburg, 79085 Freiburg,
E-Mail: christian.mair@anglistik.uni-freiburg.de

Open Access. © 2018 Christian Mair, publiziert von De Gruyter. Dieses Werk ist lizenziert unter der Creative Commons Attribution 4.0 Lizenz.
https://doi.org/10.1515/9783110538649-002

widmete, um seine Grammatik des Englischen (Fries 1952) empirisch zu fundieren, blieben die Ausnahme. Fries' Korpus umfasste ca. 250.000 Wörter. Etwas später fasste Randolph Quirk den für damalige Verhältnisse äußerst mutigen Plan, ein eine Million Wörter umfassendes Korpus des britischen Englisch zu erstellen, wobei er außer Tonbandgerät und Schreibmaschine anfangs an keine weiteren technischen Hilfsmittel dachte. Das *Survey of English-Usage*-Korpus wurde allerdings bereits während seiner Fertigstellung in den 1970er Jahren teilweise digitalisiert (Svartvik & Quirk 1980) und ist damit gleichzeitig der letzte Vertreter des korpuslinguistischen Mittelalters und eine Pionierleistung der digitalen korpuslinguistischen Neuzeit.

Bekanntermaßen hat sich die Korpuslinguistik seit diesen Tagen nicht nur in der Anglistik, sondern auch in den anderen Philologien einen zentralen Platz erobert. In Abschnitt 2 werde ich die wesentlichen Etappen einer Entwicklung rekapitulieren, durch die Ideen und Werkzeuge, die von einem kleinen Grüppchen zum Teil beargwöhnter, zum Teil belächelter Exzentriker am Rande ihrer jeweiligen Fächer entwickelt wurden, innerhalb von nur drei Jahrzehnten im Zentrum der Disziplinen angekommen sind. Abschnitt 3 wird sich einigen Schattenseiten dieser Erfolgsgeschichte zuwenden, während der abschließende Abschnitt 4 die Rolle der Technik und des menschlichen Faktors bei der Gestaltung des wissenschaftlichen Fortschritts in der Linguistik abwägt. Der Beitrag schließt mit der Einschätzung, dass sich die Korpuslinguistik gerade wegen ihres Erfolges in den nächsten Jahren im Rahmen der allgemeinen Entwicklung im Bereich der Digitalisierung in den Geistes- und Sozialwissenschaften (Stichwort *Digital Humanities*) neu positionieren muss.

2 *Oh Pioneers!* – Vom Rand in den Mainstream

Als Geburtsjahr der modernen anglistischen Korpuslinguistik gilt gemeinhin das Jahr 1964, in dem W. Nelson Francis, Henry Kučera und ihr kleines Team das Brown Corpus – offiziell „A Standard Corpus of Present-Day Edited American English, for Use with Digital Computers" – fertigstellten. Das Korpus eignete sich in erster Linie als Datengrundlage für gebrauchsbasierte Ansätze zur Beschreibung der amerikanischen Standardvarietät des Englischen, weshalb es in einer Phase, in der die frühe Generative Linguistik Chomskys und die Soziolinguistik Labovs die Sprachwissenschaft in den USA dominierten, nicht gerade auf enthusiastische Rezeption stieß. Für Generativisten waren Korpusdaten überflüssig, weil sie die wahren Herausforderungen für die wissenschaftliche Sprachbeschreibung jenseits der Sammlung relativ willkürlich ausgewählter und strukturell in den meisten Fällen denkbar unauffälliger Performanzdaten



From left: Jostein Hauge, Director of the Norwegian Computing Centre for the Humanities, W. Nelson Francis, Geoffrey Leech, Stig Johansson, Arne Zettersten, Henry Kučera, Randolph Quirk, Jan Svartvik.

Abb. 1.1: Bei der ersten ICAME-Tagung in Bergen (1979), Quelle: http://clu.uni.no/icame/history/First_ICAME_conference_Bergen_1979.gif (letzter Zugriff: 26. 10. 2017).

vermuteten. Für Soziolinguisten wiederum war der schriftsprachliche Standard die denkbar uninteressanteste Varietät, weil hier neben etwas stilistischer Variation nichts zu holen war.

Weitere Initiativen, etwa die offensichtlich sinnvolle Erweiterung der Datengrundlage um ein britisches Parallelkorpus (das LOB [= Lancaster/Oslo-Bergen]-Korpus), kamen nur schleppend voran, so dass die Gründung von ICAME (International Computer Archive of Modern English) im Jahre 1977 in Oslo eher einen ungedeckten Scheck auf die Zukunft denn die Koordination intensiver laufender Forschungsaktivitäten darstellte. Ein Foto von der ersten ICAME Tagung 1979 in Bergen zeigt eine noch eher familiäre Atmosphäre (siehe Abb. 1.1).

Wäre auf dem Bild auch noch John Sinclair zu sehen, hätte man die Pioniere der damaligen anglistischen Korpuslinguistik fast komplett versammelt. Gewisse Hinweise auf den Status und die Reputation des frühen John Sinclair lassen sich allerdings aus der zeitgenössischen Belletristik entnehmen. Dem Vernehmen nach war er die reale Inspiration für den Sprachtechnologen Robin Dempsey aus David Lodges Roman *Small World*, der seine Forschungsprogram-

matik einem staunenden Besucher mit einer Mischung aus Frankensteinschem Fanatismus und missionarischem Eifer vorträgt:

> "I'd like to take you over to our Computer Centre this afternoon," he said. "We've got something set up for you that I think you'll find interesting." He was sort of twitching in his seat with excitement as he said it, like a kid who can't wait to unwrap his Christmas presents. [...] "Anyway," he went on, "when we heard that the University was going to give you an honorary degree, we decided to make yours the first complete corpus in our tape archive." "What does that mean?" I said. "It means," he said, holding up a flat metal canister rather like the sort you keep film spools in, "It means that every word you've ever published is in here." His eyes gleamed with a kind of manic glee, like he was Frankenstein, or some kind of wizard [...] "What's the use of this?" I asked. "What's the use of it?" he said, laughing hysterically, "What's the *use*? Let's show him, Josh." And he passed the canister to the other guy, who takes out a spool of tape and fits it to one of the machines. [...] "What's your favourite word?" "My favourite word? I don't have one." "Oh yes, you do!" he said. "The word you use most frequently." "That's probably *the* or *a* or *and*," I said. He shook his head impatiently. "We instruct the computer to ignore what we call grammatical words – articles, prepositions, pronouns, modal verbs, which have a high frequency rating in all discourse. Then we get to the real nitty-gritty, what we call the lexical words, the words that carry a distinctive semantic content. Words like *love* or *dark* or *heart* or *God*. Let's see." (David Lodge, *Small world*, London: Vintage Books, 2011 [1984], p. 183 f.)

Heutige studentische Leser erkennen die Grundkonstellation: Wortsuche in digitalisierten Texten. Die Details jedoch verwundern. In ein Rechenzentrum gehen zu müssen, um eine triviale Wortsuche in einem doch recht kleinen Korpus, dem publizierten Gesamtwerk eines einzigen Autors, durchzuführen? Datenspeicherung auf filmrollenähnlichem Träger? Wozu der Techniker?

Dieser kurze Abschnitt aus einem Roman macht die beeindruckenden technischen Fortschritte der Korpuslinguistik deutlich. Viele der diesbezüglichen Probleme, mit denen sich Dempsey und sein Techniker noch abmühen mussten, sind heute gelöst. Zumindest für schriftsprachliche Daten ist die Speicherkapazität der Rechner seit Jahrzehnten kein Problem mehr – obwohl die Daten im Gegensatz zum reinen Text der frühen Korpora heute im Regelfall in XML-kompatiblem Format und mit Wortartenannotation vorliegen. Der Zugang zu den meisten Korpora ist niederschwellig möglich – vom Schreibtisch bzw. vom mobilen Endgerät des Nutzers. Die Digitalisierung von Texten ist in der Romanszene noch eine Pioniertat. Was die digitale Erfassung der Textproduktion betrifft, wäre hingegen heute eher die Frage zu stellen, wessen publiziertes Gesamtwerk noch *nicht* digital abrufbar ist.

Auch was Suchabfragen und statistische Analyse der Daten betrifft, sind die Fortschritte enorm. Suchabfrage in digitalen Textdaten ist heute kein Spezialwissen einer kleinen Avantgarde von Experten mehr, sondern essenzielle Kulturtechnik, ohne die niemand bestehen kann, der sich als Teil der moder-

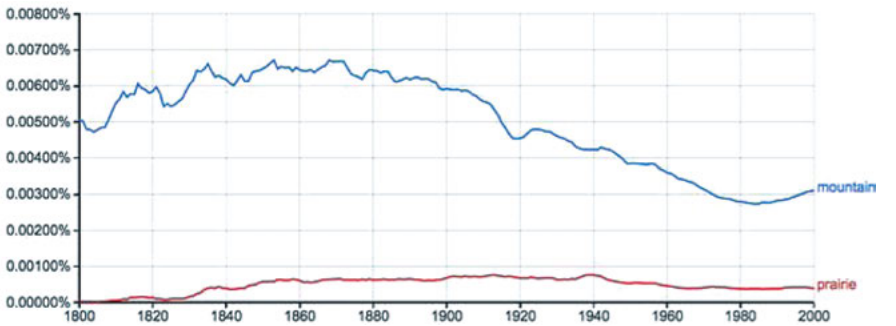


Abb. 1.2: Ngram Viewer-Diagramm „*mountain*, *prairie*“, Quelle: http://mentalfloss.com/article/60033/experiments-ngram-art (letzter Zugriff: 17. 10. 2017).

nen Welt versteht. In diesem Sinn finden sich Dempseys Nachfolger heute nicht nur in der Gemeinde der Korpuslinguistinnen und Korpuslinguisten, sondern weit darüber hinaus. Dazu gehört, wer zum Beispiel in harmloser Absicht eine Suchmaske für Informationsabfragen in Datenbanken nutzt, aber auch der wachsende Kreis derjenigen, die im Auftrag diverser Geheimdienste bei der flächendeckenden Überwachung der globalen Kommunikation tätig sind. Letzterer Bereich ist wohl derjenige, in dem in Folge des enormen Datenvolumens tatsächlich noch nicht dezentral auf dem PC gearbeitet werden kann, sondern nach wie vor riesige Rechenzentren notwendig sind. Wenn man will, einer der wenigen Fälle, wo das Gerede von „Big Data in den Geisteswissenschaften“ heute nach wie vor zutrifft.[2]

W. Nelson Francis, Geoffrey Leech, Randolph Quirk und John Sinclair hätten vermutlich auch nicht damit gerechnet, dass – fern der geheimnisumwitterten Welt der Nachrichtendienste – Korpuslinguistik zur kreativen Freizeitbeschäftigung wird – zu erkunden auf der Webseite „Experiments in Ngram Art“ von Arika Okrent (http://mentalfloss.com/article/60033/experiments-ngram-art [letzter Zugriff: 17. 10. 2017]), die den Google Ngram Viewer nützt, um Diagramme zu erstellen, die auf ikonische Weise die Bedeutung der im Google Books-Material gesuchten Wörter repräsentieren, so etwa im Falle des *mountain*, der sich über der *prairie* erhebt (siehe Abb. 1.2).

2 In diesem Zusammenhang erinnere ich an einen Artikel in der *New York Review of Books* aus dem Jahre 2014, in dem Michael Hayden, der ehemalige Chef der amerikanischen National Security Agency (NSA), mit der denkwürdigen Einsicht zitiert wird, dass „we kill people based on metadata“ (Cole 2014).

Doch zurück in den Bereich der Sprachwissenschaft: Korpuslinguistische Verfahren haben innerhalb kürzester Zeit die Lexikologie wie auch die praktische Lexikographie revolutioniert. In der Grammatikforschung waren sie eine wesentliche Voraussetzung für die aktuelle Konjunktur gebrauchsbasierter (*usage-based*) theoretischer Modelle. Weitere wichtige Impulse gab es für die Sprachgeschichte ebenso wie für Textlinguistik und Diskursanalyse. Natürlich erliegt die eine oder andere Studie der Versuchung zum statistischen Positivismus: durchzählen, was sich zählen lässt, und dann sehen wir weiter …

Intelligente Nutzung der Technik liegt dagegen dann vor, wenn Fragen gestellt werden, die man früher vermieden hätte, weil man keine Hoffnung auf Antwort gehabt hätte. So bedauert der große niederländische Sprachhistoriker Visser noch 1973:

> Today *begin* + form in *-ing* is used with striking frequency alongside of *begin* + infinitive. Which of the two alternatives predominates cannot be ascertained because of the lack of statistical data. (Visser 1970–73, III: 1888)

In der Tat ist die manuelle Auszählung von Gerundien und Infinitiven nach dem Verb *begin* in 20 Millionen Wörtern Text (um eine Zahl zu nennen, die wahrscheinlich der Jahresleistung eines geübten Lesers nahekommt) eine Aufgabe, die man keinem Mitmenschen, auch nicht einem Doktoranden, zumuten möchte. In lemmatisierten und mit Wortartenannotation versehenen Großkorpora wie dem *Corpus of Contemporary American English* (COCA) (Davies 2008–) und dem *Corpus of Historical American English* (COHA) (Davies 2010), mit zusammen fast einer Milliarde Wörtern, kann man innerhalb einiger Stunden ein Vielfaches dieser Textmenge verarbeiten und damit ein recht verlässliches Profil des Sprachgebrauchs erstellen, differenziert nach syntaktischen Umgebungen und Textsorten, und durch Heranziehung weiterer Korpora, natürlich auch nach Varietäten.

Der durchschlagende Erfolg korpuslinguistischer Methodik zeigt sich besonders deutlich, wenn sie von anderen Disziplinen zur Verfolgung eigener Ziele adaptiert wird. So haben Franco Moretti und Mark Algee-Hewitt das Stanford Literary Lab (https://litlab.stanford.edu/[letzter Zugriff: 17. 10. 2017]) aufgebaut, in dem zahlreiche kulturhistorische und literaturwissenschaftliche Fragestellungen nach den Regeln der aktuellen Korpuslinguistik und Texttechnologie erforscht werden. Beispiele für ähnliche Initiativen finden sich natürlich auch in anderen textbezogenen Wissenschaftsdisziplinen. Drei britische Historiker, Clive Emsley, Tim Hitchcock und Robert Shoemaker, haben die Archive des Londoner Zentralgerichtshofs für Strafsachen („Old Bailey“) digitalisiert und stellen ihr Projekt wie folgt vor:

> The Proceedings of the Old Bailey, 1674–1913
> A fully searchable edition of the largest body of texts detailing the lives of non-elite people ever published, containing 197,745 criminal trials held at London's central criminal court. (https://www.oldbaileyonline.org/index.jsp [Verson 7.2, letzter Zugriff: 17. 10. 2017])

Die Suchmaske dieses Korpus kommt sprachhistorisch interessierten Nutzerinnen und Nutzern nicht entgegen. Sie hilft eher bei der Frage, für welche Art von Verbrechen man zwischen 1788 und 1820 besonders leicht mit Verbannung nach Australien bestraft wurde oder ob sich zwischen Anfang und Ende des 18. Jahrhunderts die Bestrafung von diebischen Hausangestellten weiblichen Geschlechts änderte. Dieses sprachhistorische Defizit ist mittlerweile durch das – von der DFG und später auch im Rahmen von CLARIN-D geförderte – Projekt *Old Bailey Corpus* von Magnus Huber in Gießen umfassend behoben (http://www1.uni-giessen.de/oldbaileycorpus/ [letzter Zugriff: 17. 10. 2017] und http://fedora.clarin-d.uni-saarland.de/oldbailey/ [letzter Zugriff: 17. 10. 2017]). Korpora fördern eine kooperative Forschungskultur, weil einzelne Nutzer auf den Ergebnissen aller anderen aufbauen können; diese Arbeitsteilung ist effizient und ebnet den Weg zu kumulativem Wissensgewinn. Das Beispiel der doppelten Aufbereitung des Old Bailey-Materials zeigt allerdings, dass diese Art von Teamgeist derzeit noch allzu oft an den Grenzen des jeweiligen Faches endet. Es wäre zu wünschen, dass unter dem interdisziplinären Dach der *Digital Humanities* die digitale Aufbereitung von Sprachdaten in Zukunft auf eine Weise erfolgt, die Anschlussmöglichkeiten in alle relevanten fachlichen Richtungen offen hält und somit ähnliche Duplizierung von Arbeit bereits im Ansatz vermeidet.

3 Die Schattenseiten des Erfolgs

Die oben beschriebene Erfolgsgeschichte hat natürlich auch ihre Schattenseiten. So gebe ich gerne zu, dass mich bislang noch jedes System der Wortartenindizierung in Korpora geärgert hat. Ich weiß, dass häufige Funktionswörter wie *the*, *of* oder *was* in englischen Korpora mit annähernd hundertprozentiger Treffsicherheit korrekt identifiziert werden. Ich zweifle auch nicht an Angaben zur durchschnittlichen Korrektheit des *part-of-speech (POS) tagging*, die sich für standardsprachliches Material im Englischen bei 98 Prozent bewegen. Ebenso weiß ich jedoch aus eigener Erfahrung, dass die entsprechenden Raten bei jedem Thema, das mich wirklich interessiert hat, deutlich unter diesem Wert lagen. Wer zur sprachwissenschaftlich durchaus reizvollen Frage arbeiten will, welche englischen *participles* auch adjektivisch gebraucht werden kön-

nen, sollte sich niemals nur auf _V* und _J*[3] verlassen, sondern zur Sicherheit auf alle Fälle auch die nicht annotierte Version des betreffenden Korpus untersuchen. Zur Probe aufs Exempel und zur Illustration der typischen Probleme möge die folgende kleine Suche nach „STOP annoying_v?g*“ im *Corpus of Contemporary American English* (COCA) dienen (also nach Formen des Lemmas *stop*, denen ein als verbales Partizip annotiertes *annoying* folgt). Bei allen fünf Belegen, die auf diese Weise gefunden wurden, war eine adjektivische Lesart von *annoying* (in der Notation dieses Korpus *annoying_JJ*) durch das folgende Wort strukturell blockiert. Sequenzen wie *annoying the people* oder *annoying these people* können im Englischen einfach nicht als Nominalphrase fungieren, weil Artikel und Demonstrativpronomina ausnahmslos *vor* dem attributiven Adjektiv *annoying* erscheinen müssen. Diese Regularität führt zu einem hohen Grad von *Präzision* (*precision*) der Korpussuche: alle Beispiele sind im Sinne der untersuchten Fragestellung brauchbar, und keines muss als irrelevant eliminiert werden. Weniger erfolgreich ist die Suche im Hinblick auf die *vollständige Erfassung* (*recall*) der relevanten Beispiele aus dem Korpus. Wiederholt man dieselbe Suche im nicht annotierten Material („STOP annoying“), findet man ein weiteres Beispiel, und zwar eines, bei dem die verbale Lesart von *annoying* nicht strukturell erzwungen ist, sondern „nur“ durch ihre semantisch-pragmatische Plausibilität nahegelegt wird: „That way, they'll soon stop annoying people who are in the process of getting themselves well.“ Der Kontext erfordert hier die verbale Lesart („Leute zu verärgern“) und schließt die adjektivische („lästige Leute“) aus; das Annotationsprogramm hingegen vergibt den Adjektiv-Tag (*annoying_JJ*), vermutlich in Analogie zu häufigen Strukturen des Typs *stop stupid people* oder *stop silly nonsense*.

Das gravierendste Defizit in der bisherigen Geschichte der Korpuslinguistik ist jedoch der Umgang mit gesprochenen Daten, wo – wie die folgenden Beispiele zeigen werden – für die nächsten 50 Jahre noch Einiges zu tun bleibt. Abbildung 1.3 zeigt einen der legendären *slips*, mit denen jedes Wort des *Survey of English Usage*-Korpus in monumentalen Zettelkästen mehrfach nach den jeweils relevanten Kategorien abgelegt wurde.

Ohne ins Detail zu gehen, ist deutlich zu erkennen, wie reich diese orthographische Transkription mit prosodischen und paralinguistischen Informationen annotiert ist – und somit innerhalb der Grenzen des Mediums Schrift der Dynamik des gesprochenen Wortes doch einigermaßen gerecht zu werden vermag. Abbildung 1.4 zeigt dieselbe Passage, diesmal in der digitalen Version des gesprochenen Survey-Korpus, die als *London-Lund-Corpus* bekannt ist.

3 Hier in der Notation des in Lancaster entwickelten CLAWS-Taggers (http://ucrel.lancs.ac.uk/claws/ [letzter Zugriff: 17. 10. 2017]).

```
                S.1.3-9
               *bro*chùre* for# so I g /díd it#g# . and /then ¯ anóther
               *one#p#  - and
          B     *  mhm  *
         (A) /Nthèn they *said# well "/now that you've done Nthése#
               and they've been a "/sò sucNcéssful#a# we'd /like you
               to do our l Nsùpēr# . /alpha:màtic#l# a or /Nsómething#
m:narrow       and /this is one of Nthése#a# that /goes mm' sídeways#
m':rhythmic    and /fróntwards# and em/bróiders# and */dárns#m'# and
               sews */bùttons on#m#
          B    *( - laughs ) yes*
         (A) -- a and I /sāid# well a# I /don't Nrèally *think# I
m:slurred      could /wrìte# -- a and this was a m sort of m#a#
               /ninety six page :bòoklet# p /you Nknów# p# about /that
               Nbìg# *-* əm I'd I'd /used to gò through# /each of the
          B    *  m  *
         (A) *processes at :hòme# *.* a I ¯don't think it will be
              *e/nough a# just to have
```

Abb. 1.3: Ausschnitt aus Text S.1.3 (spontanes Gespräch, *Survey of English Usage*-Korpus, Svartvik & Quirk 1980: 16).

```
        BRO☆CHÙRE☆ for■ 139 so I ‖DÍD it■ · 140 and ‖then ANÓTHER one■ -
        141 and
  b     142 ☆[mhm]☆
> A     141 ‖THÈN they ▷said■ 143 well ‖now that you've done THÉSE■ 144 and
        they've been ‖SÒ SUCCÉSSFUL■ 145 we'd ‖like you to do our SÙPER■ ·
        146 ‖ALPHA▵MÀTIC■ 147 or ‖SÓMETHING■ 148 and ‖this is one of THÉSE■
        149 that ‖goes SÍDEWAYS■ 150 and ‖FRÓNTWARDS■ 151 and EM‖BRÓIDERS■
        152 and ☆‖DÁRNS■ 153 and sews☆ ‖BÙTTONS on■
  b     154 ☆( - laughs) yes☆
> A     155 - - and I ‖SĀID■ 156 well I ‖don't RÈALLY ▷think■ 157 I could ‖WRÌTE■ -
        - 158 and this was a sort of ‖ninety-six page ▵BÒOKLET■ 159 ‖you KNÓW■
        160 about ‖that BÌG■ ☆-☆ 161 [əm] I'd I'd ‖need to GÒ through■ 162 ‖each of
        the
  b     163 ☆[m]☆
> A     162 processes at ▵HÒME■ ☆·☆ 164 I don't think it will be e‖nough just to
        have
```

Abb. 1.4: Ausschnitt aus Text S.1.3 (spontanes Gespräch, *London-Lund-Corpus* (LLC), Svartvik & Quirk 1980: 85).

Man beachte, dass die 1980 als Buch publizierte digitale Version des Korpus nicht ein direkter Ausdruck des Computer-Texts ist, sondern einige redaktionelle Konzessionen in Richtung bessere Lesbarkeit macht, zum Beispiel durch

Hochstellung der Intonationspfeile von der Grundlinie. Zusätzlich eingeführt ist eine (sprachwissenschaftlich bedeutungslose) laufende Durchnummerierung der Tongruppen. Eliminiert sind dagegen zahlreiche detailliertere prosodische sowie fast alle paralinguistischen Informationen (Stimmqualität etc.). Neuere Standardkorpora wie das COCA bieten zwar wesentlich mehr transkribiertes gesprochenes Material, doch meist aus Sekundärquellen wie Transkripten, die von Rundfunk- und Fernsehanstalten mit Hilfe von linguistisch meist nicht ausgebildetem Personal erstellt werden. Hier als Beispiel ein Interview des Senders ABC mit dem bekannten Country-Musiker Tim McGraw:

(1) TIM-MCGRAW-1SINGE# We met in 1996. It was my first headlining tour and she was my first opening act for my headline tour. And we met then, got married later that year. And we toured again in 2000 for the first Soul2Soul tour. ROBIN-ROBERTS-1AB# (Voiceover) Mm-hmm. TIM-MCGRAW-1SINGE# And last year, we did Soul2Soul II. And, and we decided to do it this year because the kids are getting older... ROBIN-ROBERTS-1AB# (Voiceover) Yeah. TIM-MCGRAW-1SINGE#... and in the summer times, they're not **gonna wanna** go out on the road anymore, so I'll **have to** go by myself in, in the near future. But we thought this might be the last chance that we got to go out together.
(COCA, 2007, ABC_GMA)

In diesem *state-of-the-art mega-corpus* fehlt jede systematische prosodische und paralinguistische Annotation. Das mag zwar die Suchbarkeit erleichtern, die in der digitalen Originalversion des oben erwähnten *London-Lund-Corpus* sehr schlecht war, weil zahlreiche Wörter durch die Markierungen des Intonationsverlaufs zerhackt wurden und das Wort *brochure* (siehe auch oben, Abb. 1.4) deshalb als *bro*ch\ure** erschien, aber natürlich auch in der Form *bro*ch/ure** und zahlreichen anderen Varianten denkbar war.[4] Andererseits schränkt die radikale Reduktion der orthographischen Transkription in COCA die Tauglichkeit des Materials für Forschungen zur gesprochenen Sprache doch deutlich ein. Im Beispiel aus COCA sind die „semi-modals" *have to*, *(have)*

4 Um lexikalische Suchen zu vereinfachen, erstellte ich in den 1990er Jahren für den Freiburger Hausgebrauch eine normalisierte Zusatzversion ohne prosodische Annotation. Vor einer weiteren Verbreitung dieser schreckte ich aus Angst vor einer möglichen negativen Reaktion von Randolph Quirk zurück, der in manchen korpustheoretischen Grundsatzfragen seine Position durchaus emotional vertreten konnte. Diese normalisierte Version wurde auf Datenträgern (*floppy disks*) und mit Konkordanzsoftware (*TACT*) verwendet, die mittlerweile beide obsolet sind und hat sich somit erledigt.

got to, *going to* und *want to* durch Fettdruck hervorgehoben. Die Wahrscheinlichkeit, dass diese verbalen Ausdrücke als lexikalische Einheiten richtig erkannt und transkribiert wurden, ist hoch. Was die phonetische Realisierung betrifft, bleiben jedoch große Unsicherheiten. Die Transkription von *going to want to* mit Hilfe der beiden konventionalisierten Nicht-Standard-Schreibungen *gonna* und *wanna* legt nahe, dass auch tatsächlich Kontraktion vorlag. Der Umkehrschluss trifft nicht zu. Aus der üblichen Schreibung für *have to* darf nicht notwendigerweise auf Abwesenheit phonetischer Kontraktion geschlossen werden. Die entsprechende Schreibweise – *hafta* – ist, im Gegensatz zu *gonna* und *wanna*, noch nicht sehr fest etabliert. Nicht immer hilft Youtube mit Videoausschnitten aus den Originalsendungen, die einem die Überprüfung des gesprochenen Textes ermöglichen würden.

Vielleicht werden wir in zehn Jahren eine Situation erreichen, in der *jedes* gesprochene Korpus multimodal zur Verfügung steht: als Video- oder Tondatei, mit maschineller Unterstützung mit einer orthographischen Transkription aligniert, die in zusätzlichen Versionen mit weiterer grammatischer und prosodischer Information angereichert ist (bzw. von den Nutzerinnen und Nutzern nach eigenen Vorstellungen angereichert werden kann).

Die eklatante Unterrepräsentation spontansprachlicher Daten in Korpora weist auf eine allgemeine Gefahr bei der Entwicklung wissenschaftlicher Infrastrukturen hin. Was sich ohne großen technischen und mit vertretbarem finanziellen Aufwand machen lässt, wird gemacht: und das war in den vergangenen Jahrzehnten die Erstellung immer größerer und immer besser annotierter standardsprachlicher Schriftkorpora. Was nicht in diese Schablone passte, wurde entweder trotzdem in sie gezwängt: man vergleiche etwa die Gewinnung der gesprochenen Daten in COCA aus dem Recycling bereits verfügbarer schriftlicher Quellen, nämlich den orthographischen Transkriptionen von Medienmaterial. Auf jeden Fall verlief die Entwicklung bei der gesprochenen Sprache und ihrer gegenstandsadäquaten korpuslinguistischen Erschließung schleppender. Solche praktischen Zwänge führen dazu, dass man Begleiterscheinungen der Grammatikalisierung in gesprochener Sprache, wie eben die phonetische Reduktion von Hilfsverben, auch heute noch entweder auf der Grundlage von großen Mengen schlecht aufbereiteter Daten oder auf der Grundlage von deutlich zu geringen Mengen qualitativ hochwertiger Daten untersuchen muss. Ein Großkorpus wie COCA enthält zwar große Mengen – in der aktuellen Fassung fast 110 Millionen Wörter – gesprochener Daten, dies jedoch in einem Zustand, der, wie oben gezeigt wurde, phonetische Reduktion nicht annähernd realistisch abbildet. Diejenigen Korpora gesprochener Sprache, die wie das *Santa Barbara corpus of spoken American English* (SBCSAE) (Du Bois et al. 2000–2005) die Anforderungen an die Qualität und Überprüfbarkeit der Tran-

skriptionen erfüllen, stammen aus den 1990er Jahren und umfassen samt und sonders weniger als eine Million Wörter.[5]

Letztendlich ist die Behandlung der gesprochenen Sprache in der Korpuslinguistik ein Beispielfall für eine allgemeinere Problematik. Die Entwicklung von Forschungsinfrastrukturen erfolgt immer im Spannungsfeld zwischen technischer Innovation und der Weiterentwicklung wissenschaftlicher Fragestellungen innerhalb der fachlichen Diskussion. Das Kräfteverhältnis der Beteiligten gestaltet sich dabei allerdings derart, dass beim Aufbau von Forschungsinfrastrukturen oft die technischen Entwickler sowie ihre privaten und öffentlichen Geldgeber die Standards definieren. In diesem Innovationsprozess haben einzelne Nutzer aus den Sprachwissenschaften immer weniger Möglichkeiten, ihre inhaltlich begründete Forschungsprogrammatik als Orientierung für die technische Entwicklung einzubringen. Im schlimmsten Fall führt dies dazu, dass sie, anstatt auf die Anpassung der Infrastruktur an ihre Bedürfnisse zu drängen, ihre eigene Forschungsprogrammatik an den Möglichkeiten der Infrastruktur ausrichten: *linguistics follows technology*. Beispiele für diese Art der Kapitulation vor der Macht des Faktischen sind vielfältig und zahlreich. Eines möchte ich hier herausgreifen und ausführlicher diskutieren.

Die Korpora der postkolonialen Varietäten des Englischen orientieren sich ebenso eindeutig wie unhinterfragt am Konstrukt der nationalen Standardvarietäten des Englischen als plurizentrischer Weltsprache. Der postkoloniale Nationalstaat fungierte bereits beim ersten derartigen Großprojekt als wichtigstes Ordnungskriterium. Das *International Corpus of English* (ICE) wurde Ende der 1980er Jahre von Sidney Greenbaum als ein Cluster von analog strukturieren nationalen Einzelkorpora konzipiert (Greenbaum & Nelson 1996). Heute liegen elf Teilkorpora vollständig vor: Australien, Kanada, Ostafrika (mit Material aus Kenia, Tansania und Malawi), Großbritannien, Indien, Irland, Jamaika, Neuseeland, Nigeria, Philippinen, Singapur.[6] Nicht viel anders ist das seit 2013 verfügbare *Corpus of Global Web-based English* (GloWbE, Davies & Fuchs 2015) strukturiert; es umfasst nationale Komponenten für Großbritannien, die Vereinigten Staaten, Irland, Kanada, Australien, Neuseeland, Indien, Pakistan, Sri Lanka, Bangladesch, Malaysia, Singapur, Philippinen, Hong Kong, Nigeria, Ghana, Kenia, Tansania, Südafrika und Jamaika.

---

5 Das Santa Barbara-Korpus bietet eine prosodisch und diskursanalytisch angereicherte orthographische Transkription mit Zugriff auf die Originaltondateien. Vgl. http://www.linguistics.ucsb.edu/research/santa-barbara-corpus (letzter Zugriff: 17. 10. 2017). Das Material umfasst etwa 250.000 Wörter.

6 Weitere Korpora – unter anderem für Fiji, Ghana, Malaysia, Sri Lanka und Trinidad und Tobago – befinden sich in Vorbereitung. Mit ICE-Scotland wird erstmals eine sprachlich eigenständige Region unterhalb der nationalstaatlichen Ebene erschlossen.

Was das überwiegend muttersprachlich gesprochene Englisch Großbritanniens, Irlands, der USA sowie der ehemaligen britischen Siedlerkolonien (Australien, Neuseeland, Kanada ohne Quebec) betrifft, ist das Konstrukt der postkolonialen nationalen Standardvarietät weitgehend unproblematisch. Unstrittig ist auch, dass sich einigermaßen stabile und endonormative nationale Standards des Englischen auf nichtmuttersprachlicher Basis in Afrika (z. B. Nigeria, Ghana, Kenia) und in einigen ehemaligen britischen Kolonien Süd- und Südostasiens (z. B. Indien, Singapur) herausgebildet haben. Wo das nationale Ordnungsprinzip jedoch offensichtlich scheitert, ist Südafrika, weil in diesem Land das muttersprachlich gesprochene Englisch der Nachkommen der britischen Siedler neben den je spezifischen Zweitsprachenvarietäten der Afrikaans-sprachigen weißen Bevölkerung, der indischen Einwanderer sowie der schwarzen Mehrheitsbevölkerung existiert. Hier wie in anderen ethnisch und sprachlich heterogenen Gemeinschaften kann man daher nur hoffen, dass man aus den Metadaten des Korpus (so überhaupt vorhanden) zumindest einen Teil der faszinierenden soziolinguistischen Vielfalt rekonstruieren kann, die sich innerhalb der Grenzen eines Nationalstaats, und innerhalb eines Korpus, findet.

Die additive Reihung von nationalen Teilkorpora suggeriert auch eine egalitäre Vielstimmigkeit, die weit von der Realität der streng hierarchisch geschichteten Plurizentrik der Weltsprache Englisch entfernt ist. Auf dem indischen Subkontinent dominiert das indische Englisch, das über diasporische Verbindungen und durch wachsende mediale Präsenz weit über die Grenzen des Landes ausstrahlt. Die transnationale Wirkung des *Bangladeshi English* ist dagegen beschränkt. Im pazifischen Raum ist Einfluss des australischen Englisch auf das neuseeländische weitaus wahrscheinlicher als Einfluss in der umgekehrten Richtung. Im weltweiten Rahmen überdacht das amerikanische Englisch, zumindest im Bereich des Vokabulars, der Phraseologie und der Grammatik, mittlerweile alle anderen Varietäten, inklusive des britischen Englisch.

Potenziell noch prägender für die Forschung erwies sich bei der Erstellung dieser Korpora der *World Englishes* eine Entscheidung für die Fiktion der Einsprachigkeit als methodisches Prinzip. ICE-Jamaica ist konzipiert als Korpus des jamaikanischen Englisch, obwohl auf der Insel besonders in der gesprochenen Kommunikation ein fließendes Kontinuum zwischen Englisch und Jamaika-Kreol vorherrscht. ICE-India präsentiert sich als englisches Korpus, obwohl Code-Switching zwischen Englisch und anderen indischen Sprachen wie Hindi, Bengali oder Tamilisch in der gesprochenen Kommunikation weit verbreitet ist. Nicht anders verhielt es sich beim philippinischen Teilkorpus (ICE-PHI). Die Projektleiterin benennt in aller Offenheit die erheblichen Probleme, die bei der Erhebung spontaner englischer Gesprächsdaten zu bewältigen waren:

> This is the largest category in the whole of the corpus and, in the Philippine setting, it was a very difficult text type category to collect. Natural conversations among family members and friends in Manila are typically conducted in Tagalog, and, depending on the educational level attained, with little or much English and/or vernacular code-mixing. For purposes of the corpus, the conversations had to be in English with minimal Tagalog insertions, and therefore the collected conversations are already less natural in that respect. The presence of the tape recorder contributed to even less naturalness. (Bautista 2004: 9)

Eine ähnliche „Vorzensur" des Datenmaterials wurde auch beim gesprochenen Textmaterial aus den Medien angewandt, wie die folgende Anmerkung zur Textkategorie „broadcast discussions" belegt:

> The discussions were lifted from television (one discussion was lifted from the radio) and therefore, in general, unless the panelists were talking at the same time, the quality of the recording and of the subsequent transcript is high. In many panel discussions on local television, the language would generally be a code-mixed one. Therefore an important requirement was to identify programs that used English as the matrix language with only minimal switching to Tagalog. (Bautista 2014: 10)

Zumindest für die *Second-Language Varieties of English* erzwingt die Korpus-Schablone somit eine künstliche Homogenisierung der mehrsprachigen Realität; die Kontaktsprachen des Englischen werden entweder bereits im Vorfeld eliminiert (weil allzu mehrsprachige Texte als ungeeignetes Datenmaterial erst gar nicht in Betracht gezogen werden) oder die mehrsprachigen Anteile fristen als *extra-corpus material* ihr Quarantäne-Dasein zwischen den beiden *tags* <indig> und </indig> (für *indigenous*). Dies führt immer dann zu Verzerrungen in der Beschreibung, wenn es spezifische lokale Manifestationen sprachlicher Informalität und stilistischer Variabilität aus der Betrachtung ausklammert.

Als Nutzer macht man andererseits auch bald die Erfahrung, dass trotz aller Bemühungen, das methodisch begründete Desiderat (oder die wirklichkeitsfremde Fiktion?) der einsprachig-englischen ICE-Korpora auch bei den Zweitsprachvarietäten durchzusetzen, doch deutliche Hinweise auf die mehrsprachige Realität erhalten bleiben – wie am folgenden Beispiel aus dem philippinischen Teilkorpus gezeigt werden kann. Mit praktisch allen anderen Varietäten des Englischen teilt das philippinische Englisch die Variation zwischen neutralem bis förmlichem *because* und der umgangssprachlichen Variante *cause*. Im gesprochenen Material des Korpus (ca. 600.000 Wörter) finden sich 2.336 Vorkommen von *because* und 374 von *cause*:

(2) In fact I still have the dust of the Acropolis in my shoes **because** I came directly from <unclear> word </unclear> <,> but one thing I learned from

our short visit to Greece the cradle of democracy and the home of ancient civilization which still influences us to this day <unclear> word </unclear> higher education we must make sure that education does not go to our heads but to our minds <indig> di ko yata [m]aintindihan yun'a </indig> (<ICE-PHI:S2B-050#5:1:A>)

(3) I you know what I find amazing in college is that **because** since we came from an <.> ex </.> from an exclusive school in Bacolod I mean I find it **'cause** it's coed so I mean you know the first day that I entered school or I attended classes it's like oh so this is this is the coed school that my other friends have been experiencing ever since (<ICE-PHI:S1A-075#21:1:B>)

Man beachte, dass Beispiel (2) einen Wechsel ins Tagalog enthält (deutsch: „ich denke, ich halte es für ein"), der nicht durch einen trivialen externen Auslöser – etwa Referenz auf lokale Flora, Fauna, Nahrungsmittel oder Bräuche – bedingt ist. Der oben formulierte programmatische Anspruch auf ein englisches Korpus kann also nur teilweise eingelöst werden.

Mit der Untersuchung der beiden Varianten *because* und *cause*, und auch durch Berücksichtigung der altertümlichen dritten englischen Option, der Konjunktion *for*, ist der Variationsraum bei dieser Variable für das philippinische Englisch noch nicht abgesteckt. Neben *because* und *cause* sind nämlich noch zwei funktional äquivalente Formen aus dem Tagalog einigermaßen regelmäßig belegt, nämlich *kasi* (mit 159 Vorkommen) und *dahil* (23 Vorkommen). *Dahil* ist die traditionelle Form, während *kasi* möglicherweise aus dem Englischen entlehnt ist. Es fällt auf, dass die 23 Vorkommen von *dahil* auf Kontexte beschränkt sind, in denen Tagalog dominiert. *Kasi* dagegen tritt sowohl in solchen Code-switches auf wie in Beispiel (4) als auch als Einzelwortentlehnung wie in Beispiel (5):

(4) <indig> Siguro </indig> come to think of it <indig> siguro <indig> in this case a hidden camera is uh is is good enough <indig> **kasi** walang ano e walang </indig> uh I don't think the the this uh <unclear> word </unclear> (<ICE-PHI:S1B-038#157:1:B>)

(5) A: Like what bad things
B: Well you know like uhm you know before <indig> **kasi** </indig> it was him and <@> Bianca </@> <indig> e </indig>
A: Yeah

B: Okay they were an item and then suddenly <@> James </@> came into the picture and started telling bad things
A: About Sam
B: About Sam to the people in <@> Bianca </@> 's dormitory
A: Oh really
(<ICE-PHI:S1A-097#45:1:A> bis <ICE-PHI:S1A-097#51:1:A>)

Beispiel (5) zeigt auf jeden Fall, dass es neben der internationalen informellen Variante *cause* im philippinischen Englisch auch noch eine spezifisch lokale Möglichkeit gibt, Informalität zu signalisieren, nämlich *kasi*. Über den Einzelfall hinaus zeigt das Beispiel, dass es verfehlt wäre, in scheinbar unordentlichen Korpusdaten aufräumen zu wollen, indem *kasi* in Beispiel (5) als Lehnwort und somit Teil des philippinischen Englisch klassifiziert, in Beispiel (4) jedoch als einen Anglizismus im Tagalog. Vielmehr geht es darum zu zeigen, dass eine vollständige Beschreibung des philippinischen Englisch nur im mehrsprachigen Kontext möglich ist und eine neue Generation von ICE-Korpora diesem Umstand in ihrem Design vielleicht offensiver Rechnung tragen könnte.

Im ICE-Korpus, wie es konzipiert wurde, ist eine Passage wie in Beispiel (6) ein Betriebsunfall, der auf der Grundlage des theoretisch formulierten Anspruchs nicht hätte passieren sollen:

(6) <indig> Ngayong hapon mga kaibigan nakita niyo po yung </indig> medal tally <indig> mukhang magiging </indig> exciting <indig> yung labanan ng US at China dahil tatlong </indig> gold medals <indig> na lang ang uh lamang ng US sa China </indig> so any time any moment <indig> ay pwedeng uh mapantayan or malampasan ng China ito dahil </indig> ongoing <indig> pa rin po ang mga </indig> gold medal matches <indig> dito po sa </indig> twenty-seventh Olympiad
(<ICE-PHI:S2A-019#3:1:A>)

In einem noch zu erstellenden mehrsprachigen *Corpus of Educated Philippine Spoken Usage* könnten auf der Grundlage solcher Daten hingegen völlig neue Einblicke in eine komplexe sprachliche Realität gewonnen werden.

Die Französische Revolution wird oft dafür verantwortlich gemacht, dass sich in ihrer Folge in Europa und in den von ihm kolonisierten Teilen der Welt eine Ideologie der Einheit von Sprache und Nation verbreitet habe. Angesichts des Überwiegens einsprachiger Korpora und der nach wie vor starken nationalen Traditionen in der Korpusforschung ließe sich durchaus die Frage stellen, ob der lange Schatten der Französischen Revolution bis in die heutigen digitalen Forschungsinfrastrukturen reicht: Eine Nation mit einer Sprache braucht

ein einsprachiges Nationalkorpus.[7] Vielleicht ist die Zeit gekommen, die Dominanz einsprachiger schriftorientierter Standard-Korpora in Frage zu stellen und die Arbeit mit Nichtstandardvarietäten, mit gesprochener Spontansprache und mit mehrsprachigen Daten als neue Forschungsprioritäten der Korpuslinguistik zu etablieren. Die technischen Entwickler und die Geldgeber könnten nach dieser Vorgabe bei der Umsetzung der neuen Forschungsagenda helfen. Dies wäre eine Umkehrung der bisher üblichen Arbeitsabläufe, bei denen sich Wissenschaftlerinnen und Wissenschaftler allzu oft an den technischen Gegebenheiten ausrichten mussten, doch würde es an innovativen, wissenschaftlich anspruchsvollen und auch gesellschaftlich nützlichen Forschungsvorhaben nicht mangeln.

4 Wissenschaftlicher Fortschritt im Spannungsfeld zwischen Mensch und Technik

Wenn wir auf die Geschichte der Linguistik zurückblicken, können wir feststellen, dass manche wissenschaftliche Revolutionen im Fach von der technischen Entwicklung weitgehend entkoppelt verliefen. So benötigte Ferdinand de Saussure für die Konzeption des strukturalistischen Ansatzes wohl nicht mehr als Papier und Bleistift. Soziolinguistik und Diskursanalyse andererseits wären ohne die Erfindung von Tonaufnahmetechniken und der damit ermöglichten Speicherung akustischer Daten wohl nicht denkbar. Und jede Verbesserung der technischen Ausstattung – z. B. Erhöhung der Aufnahme- und Wiedergabequalität oder Verkleinerung der Geräte für den mobilen Einsatz – erweitert sofort den Horizont für die wissenschaftliche Arbeit. Michael Halliday hat insofern nur wenig übertrieben, wenn er feststellt:

> Perhaps the greatest single event in the history of linguistics was the invention of the tape recorder, which for the first time has captured natural conversation and made it accessible to systematic study. (Halliday 1994: xxiii)

Einen ähnlichen Innovationsschub verdankt die Linguistik auch dem Computer. Einerseits funktioniert all das, was man mit Tonband und Videorekorder machen kann, mit digitalen Verfahren noch besser und schneller. Andererseits

7 Das *British National Corpus* (BNC) ist in dieser Hinsicht eine Fehlbenennung, weil die englische Sprache auf so offensichtliche Weise größer ist als die Summe der Nationalstaaten, in denen sie als Muttersprache gesprochen wird. Das *Russian National Corpus* ist ein Grenzfall. Das ungarische, polnische und tschechische Nationalkorpus entsprechen dem Programm.

entwickelten sich – zumindest für die kleine Zahl der Sprachen, die traditionell die größte wissenschaftliche Aufmerksamkeit auf sich ziehen – sehr rasch luxuriöse korpuslinguistische Arbeitsumgebungen, die an Fülle und Diversität des Datenmaterials alles Frühere bei Weitem in den Schatten stellten.[8]

Resultat war ein ungeahnter Aufschwung von gebrauchsbasierten Beschreibungsansätzen und statistischen Analyseverfahren. Dort, wo sich technische und konzeptuelle Innovation im Gleichgewicht befinden, herrschen fast paradiesische Zustände. Bei vielen Fragen haben sich die Mühen, die vor der Digitalisierung mit der Datenbeschaffung und -aufbereitung verbunden waren, drastisch verringert. „First throw away your evidence!" war der Slogan, mit dem John Sinclair (1986) die neue linguistische Arbeitspraxis plakativ beschrieb. Korpora ermöglichen neue datengetriebene (*bottom up*) Ansätze und erhöhen das Potenzial und die Effizienz etablierter konzeptuell-theoretisch motivierter (*top down*) Verfahren. Einen guten Beleg dafür liefert die Vielzahl jüngerer korpuslinguistischer Arbeiten zur Grammatikalisierung, in der ein über hundert Jahre alter theoretischer Begriff in einem neuen technischen Umfeld wieder zum Strahlen gebracht wird (vgl. den Forschungsüberblick in Mair 2011).

Ein letzter Aspekt sollte bei der Bilanz nicht vergessen werden. Quer über alle Einzelthemen hinweg führen Korpora auch zu einer neuen Arbeitskultur in der Sprachwissenschaft. Sie fördern kooperative Forschung und kumulativen Erkenntnisgewinn, weil viele Wissenschaftlerinnen und Wissenschaftler von einer gemeinsamen Datengrundlage ausgehen. Was die Zukunft des Fachs betrifft, muss klar sein, dass gerade die mehrsprachigen Korpora, deren Erstellung oben als Desiderat formuliert wurde, nur in Teamarbeit kompiliert und analysiert werden können. Korpora haben auch das Potenzial, gewisse (sprach)wissenschaftliche Hierarchien zu unterminieren, so etwa diejenige zwischen Linguisten als *native speaker* und *non-native speaker* der untersuchten Sprache (wovon ich als deutschsprachiger Anglist selber profitiert habe). Und schließlich gelangt der Nachwuchs in der Korpuslinguistik oft sehr früh zu Autorität: durch Technikaffinität und Vertrautheit mit komplexen Verfahren der statistischen Analyse und Visualisierung.

8 Ich erinnere an die Einschränkung, die ich oben für spontansprachliche Daten gemacht habe.

5 Die Selbstauflösung der Korpuslinguistik und ihre Neukonzeption als Teil der *Digital Humanities*

Nach einer fünfzigjährigen stürmischen Entwicklung mit insgesamt sehr positiver Bilanz zeichnet sich für die Korpuslinguistik der nächste Umbruch ab. Bis in die frühen 1990er Jahre benutzten nur sehr wenige Sprachwissenschaftlerinnen und Sprachwissenschaftler digitale Korpora, und die Mehrheit der *Community* hatte wenig Ahnung, wozu solche Werkzeuge gut sein sollten und wie sie im Detail funktionierten. Diese Situation hat sich grundlegend gewandelt. Korpora sind überall. Der Einstieg in die Korpuslinguistik ist niederschwellig möglich. Der prinzipielle Nutzen der neuen Verfahren wird nicht in Frage gestellt. Gerade wegen des Überangebots und der raschen technischen Entwicklung ist es heute für Einsteiger jedoch schwieriger geworden, ein Verständnis dafür zu entwickeln, wie die Werkzeuge im Detail funktionieren und was ihre Potenziale und Grenzen sind.

In gewisser Weise ist der Korpuslinguistik sogar ein zentrales Identifikationsmerkmal abhandengekommen, nämlich der Konsens darüber, was ein Korpus ist. Die enge Definition des linguistischen Korpus, als digitale Textsammlung, die von Sprachwissenschaftlern für die Zwecke sprachwissenschaftlicher Analyse erstellt wurde, gilt nur mehr bedingt. Viele Korpora werden nicht mehr nach traditionellen Verfahren kompiliert und dann digitalisiert, sondern durch Recycling bestehender Digitalisate erzeugt. Nicht nur die Linguistik, sondern alle Wissenschaften, die mit sprachlichen Daten arbeiten, operieren in einem Universum von digitalisierten Texten, in das zunehmend auch multimodale Sprachdaten aufgenommen werden. Um mit dem Bild zu sprechen, das Gatto (2011) geprägt hat: Die feste Grundlage eines klar definierten Korpus („body") ist verloren gegangen, und an ihre Stelle ist ein vielfältiges, reiches, aber auch verwirrend unübersichtliches digitales Sprachnetz getreten („web").

„Einführungen in die Korpuslinguistik" für Studierende im Grundstudium[9] wird und soll es auch in Zukunft geben. Eines ihrer wichtigen Lehrziele wird jedoch sein, ein Bewusstsein dafür zu schaffen, dass die interessantesten digitalisierten Sprachdaten immer öfter nicht in Form traditioneller Korpora vorliegen werden und dass die Linguistik ihre vormalige Führungsrolle bei der krea-

9 Vgl. z. B. Gerstenberg (2013) für die Romanistik, Lemnitzer & Zinsmeister (2010) bzw. Perkuhn, Keibel & Kupietz (2012) für die Germanistik und McEnery & Hardie (2012) für die Anglistik.

tiven wissenschaftlichen Arbeit mit Korpora und digitalen Sprachdaten zunehmend mit anderen Disziplinen teilen wird müssen. Am Ende einer beeindruckenden Erfolgsgeschichte ist die Korpuslinguistik somit dabei, sich aufzulösen und in die *Digital Humanities*-Bewegung zu integrieren.

Literatur

Bartlett, John (1913): *A complete concordance or verbal index to words, phrases, and passages in the dramatic works of Shakespeare*. London: Macmillan.

Bautista, Ma. Lourdes (2004): An overview of the Philippine component of the International Corpus of English. *Asian Englishes* 7, 8–26.

Cole, David (2014): We kill people based on metadata. *New York Review of Books*, 10. Mai 2014. http://www.nybooks.com/daily/2014/05/10/we-kill-people-based-metadata/ (letzter Zugriff: 17. 10. 2017).

Davies, Mark (2008–): *The Corpus of Contemporary American English (COCA): 520 million words, 1990–present*. https://corpus.byu.edu/coca/ (letzter Zugriff: 17. 10. 2017).

Davies, Mark (2010–): The Corpus of Historical American English (COHA): 400 million words, 1810–2009. https://corpus.byu.edu/coha/ (letzter Zugriff: 17. 10. 2017).

Davies, Mark & Robert Fuchs (2015): Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-Based English Corpus (GloWbE). *English World-Wide* 36, 1–28.

Du Bois, John W., Wallace L. Chafe, Charles Meyer, Sandra A. Thompson, Robert Englebretson & Nii Martey (2000–2005): *Santa Barbara corpus of spoken American English, Parts 1–4*. Philadelphia, PA: Linguistic Data Consortium.

Fries, Charles C. (1952): *The structure of English: The construction of English sentences*. New York: Harcourt, Brace.

Gatto, Maristella (2011): The 'body' and the 'web': The web as corpus ten years on. *ICAME Journal* 35, 35–58.

Gerstenberg, Annette (2013): *Arbeitstechniken für Romanisten*. Berlin, Boston: de Gruyter Mouton.

Greenbaum, Sidney & Gerald Nelson (1996): The International Corpus of English (ICE) project. *World Englishes* 15, 3–15.

Halliday, M. A. K. (1994*): An introduction to functional grammar*. 2. Aufl. London: Edward Arnold.

Lemnitzer, Lothar & Heike Zinsmeister (2010): *Korpuslinguistik: Eine Einführung*. Tübingen: Narr.

Mair, Christian (2011): Grammaticalization and corpus linguistics. In Heiko Narrog & Bernd Heine (Hrsg.), *The Oxford handbook of grammaticalization*, 239–250. Oxford: Oxford University Press.

McEnery, Tony & Andrew Hardie (2012): *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

Perkuhn, Rainer, Holger Keibel & Marc Kupietz (2012*): Korpuslinguistik*. Paderborn: Fink.

Sinclair, John (1986): First throw away your evidence! In Gerhard Leitner (Hrsg.), *The English reference grammar: Language and linguistics, writers and readers*, 56–64. Tübingen: Niemeyer.

Svartvik, Jan & Randolph Quirk (Hrsg.) (1980): *A corpus of English conversation* (Lund Studies in English 56). Lund: Liber/Gleerups.
Visser, Frederikus Th. (1970–1978): *An historical syntax of the English language*. 3 Bde. Leiden: Brill.