

# Einleitung

**Daniel ELMIGER & Alain KAMBER**

Universität de Genève & Universität de Neuchâtel

Die Korpuslinguistik hat sich in den vergangenen Jahren einerseits zu einem eigenständigen Forschungsgebiet entwickelt, andererseits sind korpuslinguistische Ansätze auch in anderen sprachwissenschaftlichen Feldern fruchtbar genutzt worden, teilweise unterstützend, als komplementäre Form der Datenerhebung und -auswertung, teilweise auch, indem methodologisches und interdisziplinäres Neuland betreten worden ist.

Für die korpuslinguistische Arbeit stehen heute sowohl eine Vielzahl von großen, bereits zugänglich gemachten Korpora zur Verfügung (von Großkorpora wie *DeReKo*, *Wortschatzportal Uni-Leipzig* oder *DWDS* bis zu einer kaum noch überschaubaren Anzahl von kleineren Spezialsammlungen) als auch viele Instrumente zur Korpuserstellung und -befragung (Tagger, Parser, Lemmatisierer, Abfrageinstrumente usw.).

Während sich die technischen Möglichkeiten immer mehr erweitern, wird auch der Fokus für den Einsatz korpuslinguistischer Ansätze immer breiter. In gewisser Weise verschwinden Grenzen zwischen rein wissenschaftlichen Verwendungszwecken und auch von einem breiten Publikum benutzbaren Anwendungen (wie im Fall der Leipziger Korpora oder Google Ngram Viewer), die sich von reinen Internetsuchen erheblich unterscheiden. Von diesen zahlreichen Möglichkeiten zeugen diverse Einführungen und Überblicksdarstellungen (McEnery & Wilson, 2001; Lemnitzer & Zinsmeister, 2006; Bubenhofer, 2009; Lüdeling & Kytö, 2008, 2009; O’Keeffe & McCarthy, 2010). Andere Werke gehen auf spezifischere Bereiche ein, wie beispielsweise Korpuspragmatik (Felder, Müller & Vogel, 2011) oder Korpus in Übersetzungswissenschaften (Oakes & Meng, 2012), andere wiederum sind enger gesteckten Themen gewidmet, wie den Satzadverbien (Duffner, 2010) oder den usuellen Wortverbindungen (Steyer, in Vorb.).

Auch im vorliegenden Band spiegeln sich diese unterschiedlichen Perspektiven wider. So wird neben eher "klassischen" Fragestellungen eine Reihe von neuen weiteren Einsatzgebieten aufgezeigt. Angeführt werden die Beiträge durch zwei Studien zum Einsatz des Web als Korpus (Bickel, Deronne) und einen Grundsatzartikel zur Messbarkeit von Daten (Quasthoff & Eckart). Es folgen drei Studien zur Stilistik mit einem vorwiegend automatisch-quantitativen Ansatz (Mayaffre, Magri-Mourgues, Savoy). Die vier nächsten befassen sich mit Fragen der Grammatik und der Lexik (Näf, Diwersy & François, Elmiger) sowie der Translatologie (Tauveron). In den

letzten drei Beiträgen werden weitere Einsatzmöglichkeiten der korpuslinguistisch basierten Methodologie erschlossen: Fremdsprachenunterricht (Kamber), Wortverbindungen und Phraseologie (Steyer), und schließlich Korpuspragmatik (Bubenhofer & Scharloth).

Zusammenfassend kann an dieser Stelle festgestellt werden, dass die jahrelang aufrecht erhaltene und bisweilen recht polemisch vertretene Unterscheidung zwischen korpusbasierter bzw. korpusgesteuerter Analyse und korpusgestützter Überprüfung dann an Schärfe verliert, wenn Korpuslinguistik in den Dienst eines anderen Ansatzes gestellt wird.

### **Das Web als Korpus**

In seinem Beitrag liefert Hans BICKEL zunächst einen Erfahrungsbericht über die Benutzung von *AltaVista* und später auch *Google* für lexikografische Zwecke, genauer gesagt für die Arbeit am *Variantenwörterbuch des Deutschen* zwischen 1998 und 2004. Während es in dieser Zeitspanne möglich war, dank solcher Suchmaschinen für die sprachwissenschaftliche Forschung verlässliche Ergebnisse zu erzielen, so hat sich die Situation in der Zwischenzeit verschlechtert, wie die Wiederholung bestimmter Vorgänge zeigt. Zudem fehlen auch heute noch – trotz mehrerer neuerer Projekte zur Erschaffung von Datenbanken – strukturierte Korpora, welche es ermöglichen würden, nationale Varianten eingehend zu analysieren.

Auch im zweiten Beitrag geht es um das Internet als mögliche Quelle für korpuslinguistische Untersuchungen: In seiner Arbeit über die syntaktische und semantische Beschreibung französischer Verben geht Emmanuel DERONNE detailliert auf die Vor- und Nachteile eines "Webkorpus" ein, das zwar einerseits leicht eine Vielzahl von – allenfalls auch seltenen – Belegen liefert. Andererseits gilt es zu bedenken, dass sich Webanfragen auf eine gewaltige Anhäufung ganz unterschiedlicher Texte und Textformen beziehen, wobei die Suchresultate auf den ersten Blick jeweils alle denselben Stellenwert zu haben scheinen. Der Autor plädiert für eine gestaffelte Suche, bei der fortlaufend Resultate generiert werden, die analysiert werden und in weitere Suchanfragen einfließen.

### **Messbarkeit und Stilistik**

In ihrem Beitrag "Vergleichbarkeit von sprachstatistischen Messungen" liefern Uwe QUASTHOFF und Thomas ECKART grundsätzliche Erwägungen zur Messbarkeit von Daten. Um verschiedene Texte, Korpora und auch Sprachen miteinander zu vergleichen, müssen Messgrößen und Messverfahren möglichst eindeutig definiert und operationalisiert werden. Anhand von Beispielen, die im Zusammenhang mit der Erarbeitung eines Frequenz-

wörterbuchs gemacht worden sind, zeigen die Autoren auf, wie die einzelnen Verarbeitungsschritte (Datenaufbereitung, Satzsegmentierung usw.) sowie gewisse Festlegungen wie die Wortdefinition oder das Verfahren zur Silbenzählung sich auf die gemessenen Ergebnisse (z. B. Wortanzahl, Type-Token-Ratio oder Satzlänge) auswirken.

Jacques SAVOY geht in seinem Beitrag "Who Wrote this Novel" der Frage nach, welche automatisierbaren Verfahren der Textanalyse dazu beitragen können, die individuelle Autorenschaft von Texten computergestützt zu bestimmen. Im Besonderen werden Verfahren berücksichtigt, die sich auf (seltene bzw. frequente) Textformen bzw. Lemmata stützen. Der Autor vergleicht zwei Verfahren zur Bestimmung der Autorenschaft und berücksichtigt dabei Daten von Autoren verschiedener Sprachen (Deutsch, Französisch und Englisch), was es erlaubt, die jeweiligen Vor- und Nachteile der getesteten Verfahren empirisch gestützt zu eruieren. Es zeigt sich, dass bei der Abgrenzung von Texten verschiedener Autoren das Z-Score-Verfahren, das sich im Besonderen auf textspezifisches Vokabular abstützt, gute Ergebnisse zeitigt.

Véronique MAGRI-MOURGUES will in ihrem Beitrag aufzeigen, wie ein zunächst rein quantitatives Analyseverfahren zur Interpretation von literarischen Texten beitragen kann, dies sowohl auf stilistischer als auch auf inhaltlicher Ebene. Mit der Software Hyperbase untersucht sie Rousseaus Werke und bringt dank der Darstellung der Ergebnisse in Form von Netzwerken markante assoziative Gedankenmuster des Schriftstellers ans Licht. Anhand einer hermeneutischen Vorgehensweise schlägt sie einen Bogen zwischen der quantitativen und der qualitativen Analyse, da die erbrachte Kookkurrenzanalyse sich vom einzelnen Satz bis zum gesamten Korpus erstreckt.

Ausgangspunkt für die Untersuchungen von Damon MAYAFFRE sind politische Texte, im Besonderen die Manuskripte von Reden französischer Politiker, in denen der Autor die Frequenz adverbialer Formen untersucht. Im Französischen lassen sich Adverbien zwar nicht trennscharf von anderen Wortarten unterscheiden; die logometrischen Analysen von Mayaffre tragen jedoch dazu bei, mit Hilfe von quantitativen Resultaten den stilistischen und pragmatischen Wert von Adverbien (bzw. Adverbialisierungen) in der politischen Rede zu bestimmen. Das der Untersuchung zugrunde liegende Korpus wird sowohl quantitativ als auch qualitativ ausgewertet, wobei stilistische Feinheiten des Gebrauchs situativ und interpretativ berücksichtigt werden können.

### **Grammatik, Lexik und Translatologie**

Anton NÄF plädiert in seinem Beitrag für eine neuinterpretierte und zeitgemäßere Darstellung der Modalverben im Deutschen. So sollte die Form

*möchten* im heutigen Deutsch nicht länger als Konjunktiv II-Form von *mögen* betrachtet werden, sondern als eigenständiges, siebtes Modalverb neben *dürfen*, *können*, *mögen*, *müssen*, *sollen* und *wollen*. Die korpuslinguistische Analyse beweist ohne Zweifel, dass die beiden Verben *mögen* und *möchten* auf morphologischer, semantischer und pragmatischer Ebene zu differenzieren sind. Dabei ermöglichen die Kookkurrenzprofile der jeweiligen Verbformen in der Kookkurrenzdatenbank *CCDB* und im *Wortschatz-Portal Uni-Leipzig* eine kritische Auseinandersetzung mit den Angaben in den deutschsprachigen Wörterbüchern.

Sascha DIWERSY und Jacques FRANÇOIS untersuchen in einer Vorstudie zu einem breit angelegten Projekt die Kombinatorik von kausativen Funktionsverben mit emotionsbezeichnenden Substantiven (z. B. {*faire & peur*} oder {*semer & doute*}). Anhand eines journalistischen Großkorpus aus dem Jahr 2007 wird die Frequenz bestimmter Kombinationen ermittelt und dargestellt. Es zeigt sich, dass von den mehreren tausend gefundenen Verbindungen nur wenige wirklich frequent sind und dass sich deren "kollokatives Gewicht" beträchtlich voneinander unterscheidet. Es scheint, dass bestimmte Faktoren (wie etwa der Numerus oder das syntaktische Umfeld der Substantive) die Häufigkeit bestimmter Kombinationen entscheidend beeinflussen.

In seinem Beitrag untersucht Daniel ELMIGER Bildung und Gebrauch von substantivierten Partizipialformen als Personenbezeichnungen. Er fokussiert dabei auf die neutralisierenden Partizip-I-Formen, die sich von jedem Verb ableiten lassen und die im Plural im Rahmen einer "nichtsexistischen Sprachverwendung" große Dienste leisten, indem sie generisch gebrauchte Maskulinformen zu vermeiden helfen. Die Untersuchung eines Ausschnitts aller öffentlich zugänglichen Korpora des Archivs der geschriebenen Sprache (DeReKo) zeigt, dass sich unter den Partizip-I-Formen in den letzten zwanzig Jahren eine hohe Anzahl von Gelegenheitsbildungen befindet und dass manche Formen darüber hinaus sich nicht einfach durch die Substantivierung eines Verbs erklären lassen.

Durch die Befragung des Parallelkorpus *Europarl* gelingt es Matthias TAUVERON, die beiden wichtigsten Übersetzungsmöglichkeiten für das französische Wort *action* im Schwedischen zu ermitteln. Die kontextbezogene Untersuchung der beiden Wörter *handling* und *åtgärd* lässt jedoch erkennen, dass sie verschiedene Kookkurrenten haben und dementsprechend semantische Unterschiede aufweisen. Die Prozedur zeigt, dass die Arbeit mit Parallelkorpora und die Berücksichtigung von kontextsensitiven Übersetzungsmöglichkeiten – vielmehr als diejenige mit einsprachigen Datenbanken – es dem Forscher erlaubt, den verschiedenen Bedeutungen eines Lexems auf die Spur zu kommen und das Phänomen der Polysemie in ein neues Licht zu rücken.

## Weitere Anwendungsmöglichkeiten

Alain KAMBER stellt in seinem Beitrag die Korpuslinguistik in den Dienst des Sprachenunterrichts (im vorliegenden Fall: Französisch als Fremdsprache), indem er Vorschläge macht, wie Studierende Großkorpora (wie das *Corpus français*) dazu benutzen können, um auf relativ unkomplizierte Weise Hypothesen über den Sprachgebrauch zu formulieren und sie direkt anhand von authentischem Datenmaterial zu verifizieren. In einem ersten praktischen Beispiel geht es um die Kookkurrenzanalyse eines bestimmten Verbs (*entourer*) und in einem zweiten Beispiel um die Bedeutungsunterscheidungen von quasisynonymen Wörtern aufgrund ihres Kontexts. Der Autor plädiert für die Benutzung von authentischem Korpusmaterial zur Förderung der Individualisierung und Autonomisierung des Unterrichts.

Um Usuelle Wortverbindungen (UWV), d. h. aus mehreren Wörtern bestehenden, mehr oder weniger syntaktisch festen Einheiten, geht es im Artikel von Kathrin STEYER. Mit Hilfe des sehr umfangreichen *Deutschen Referenzkorpus* zeigt sie auf, wie induktive, datengeleitete Analyseverfahren dazu beitragen können, die Beschreibung verfestigter Sprachgebrauchsmuster zu verbessern. Anhaltspunkt für die Analyse sind nicht syntaktische Strukturen (z. B. Nominal- oder Verbalphrasen) oder A-Priori-Annahmen, sondern Regularitäten, die auf der Textoberfläche festgestellt werden können und deren Musterhaftigkeit in einem mehrschrittigen, iterativen Vorgehen aufgezeigt wird. Dabei zeigt sich, dass der Gegenstandsbereich der Phraseologie weniger lexikalisiert ist, als dies bisherige Beschreibungen nahelegen.

In ihrem Beitrag "Korpuspragmatische Analysen alpinistischer Literatur" loten Noah BUBENHOFER & Joachim SCHARLOTH die Grenzen der Korpuslinguistik in Richtung Soziolinguistik, Diskurslinguistik und Kulturanalyse aus. Durch den datengeleiteten, korpuslinguistischen Zugang gelingt es ihnen, wiederkehrende sprachliche Muster ("Sprachgebrauchsmuster") als Ergebnis von sprachlich-sozialem Handeln zu verstehen und zu deuten. Als diachronisches Korpus verwenden sie dabei die von 1864 bis 2009 erschienenen Ausgaben des Jahrbuchs des Schweizer Alpenclubs. Die Analyse der auftretenden formalen Veränderungen im Laufe der Jahre darf zumindest teilweise als Produkt veränderter Denk- und Handlungsweisen gelesen werden.

## Ehre, wem Ehre gebührt

Ein Teil der in diesem Band zusammengefassten Arbeiten sind im Zusammenhang mit einer Tagung zur Korpuslinguistik entstanden, die am 13. Mai 2011 an der Universität Neuenburg (Université de Neuchâtel) statt-

gefunden hat<sup>1</sup>. Am Deutschen Institut dieser Universität hat Anton Näf, dem diese Tagung gewidmet war, bis zu seiner Emeritierung als Professor für Mediävistik und deutsche Sprachwissenschaft – eine wohl im Aussterben begriffene Doppelspezialisierung – gelehrt und geforscht. Obwohl er sich im Laufe seiner Karriere ebenfalls für andere Gebiete interessiert hat, wie Althochdeutsch, Schulsprachenpolitik, Fremdsprachenlernen usw., hat Anton Näf als einer der ersten in der Schweiz überhaupt begriffen, welche Chance sich für die Sprachwissenschaft bietet, anhand von digitalen Großkorpora und entsprechenden Abfragewerkzeugen mit der Erforschung der sprachlichen Fakten sozusagen nochmals "von vorn" zu beginnen.

Angereichert wurden die Beiträge der Tagung durch eine Reihe von weiteren Artikeln, von denen die meisten von französischsprachigen Kolleginnen und Kollegen stammen. Auf diese Weise ist im vorliegenden TRANEL-Band ein Zeugnis der mehrsprachigen Forschung entstanden, wie sie nicht nur für die Schweiz bezeichnend ist, sondern auch für die Sprachwissenschaft an der Universität Neuenburg (wenige Kilometer von der Sprachgrenze entfernt) und für das langjährige und vielsprachige Wirken Anton Näfs.

## Literatur

- Bubenhofer, N. (2009): Sprachgebrauchsmuster. Korpuslinguistik als Methode der Diskurs- und Kulturanalyse. Berlin; New York: Walter de Gruyter
- Duffner, R. (2010): Die Satzadverbien im Deutschen: eine korpusbasierte Untersuchung. Dissertation. Online abrufbar unter <http://doc.rero.ch/record/25013>. Université de Neuchâtel.
- Felder, E., Müller, M. & Vogel, F. (Hg.) (2011): Korpuspragmatik. Thematische Korpora als Basis diskurslinguistischer Analysen. Berlin; New York: De Gruyter
- Lemnitzer, L. & Zinsmeister, H. (2006): Korpuslinguistik. Tübingen: Gunter Narr Verlag.
- Lüdeling, A. & Kytö, M. (ed.) (2008; 2009): Corpus Linguistics. An International Handbook. Berlin; New York: Walter de Gruyter (= HSK; 35.1)
- McEnery, T. & Wilson, A. (2001): Corpus linguistics: an introduction. Edinburgh: Edinburgh University.
- Oakes, M. P. & Meng J. (ed.) (2012): Quantitative Methods in Corpus-Based Translation Studies. A practical guide to descriptive translation research. Amsterdam et al.: John Benjamins (= Studies in Corpus Linguistics; 51).
- O'Keeffe, A. & McCarthy, M. (2010): The Routledge Handbook of Corpus Linguistics. London; New York: Routledge.
- Steyer, K. (in Vorb.): Usuelle Wortverbindungen. Zentrale Muster des Sprachgebrauchs aus korpusanalytischer Sicht. Tübingen: Gunter Narr Verlag (Studien zur deutschen Sprache).

---

<sup>1</sup> Organisiert wurde diese Tagung neben den beiden Unterzeichneten von unseren Kolleginnen und Kollegen Rolf Duffner, Natacha Reynaud Oudot, Gabrielle Schmid und Eva Wieden Keller.