Łukasz Rogoziński, Grzegorz Zarzeczny, Marta Rogozińska
Wrocław, Polen

# Aufbereitung und Erstellung eines Fachkorpus der gesprochenen Sprache (am Beispiel des polnischen Teils des GeWiss-Korpus)

## 1. Einführung

Ein natürlicher Ausgangspunkt für die Forschungen aus dem Bereich der Korpuslinguistik ist die Aufbereitung eines entsprechenden Korpus von Texten. Das Hauptproblem auf dieser Arbeitsetappe ist die Repräsentativität des Korpus: Die Authentizität der Daten, der angemessene Korpusaufbau, das Gleichgewicht zwischen den Einzelteilen (in der Regel der Genres) etc. (vgl. Köhler 2005:5-7). Im Falle der Korpuserstellung zu Vergleichszwekken bei mehrsprachigen Texten muss man unbedingt auch die kulturellen Unterschiede mitberücksichtigen, z. B. im Bereich der scheinbar parallelen Genres. Letztendlich erfordert der Aufbau eines Korpus der gesprochenen Sprache auch eine Vereinheitlichung von technischer Seite her – eine vergleichende Qualität der Audio- und Videoaufnahmen, vergleichende Grenzbedingungen einer zugelassenen Aufnahme etc. Die Arbeitsetappe hat in einem gewissen Maße einen apriorischen und deduktiven Charakter.

Der vorliegende Beitrag thematisiert die weitere Arbeitsetappe, die zu der vorherigen im Kontrast steht: die Erstellung eines Fachkorpus von gesprochenen Texten. Selbst wenn ein Korpus so präzise wie nur möglich entworfen wird, wird seine endgültige Gestalt von vielen Faktoren abhängen, die zu Beginn der Korpusarbeiten nicht voraussehbar sind. Dieses Problem bezieht sich viel mehr auf die Korpora der gesprochenen Sprache als die der geschriebenen. Es scheint jedoch, dass diese Tatsache in den korpusbasierten Arbeiten sehr oft ausgelassen wird. Den Ausgangspunkt für die weiteren Überlegungen bilden die Erfahrungen bei der Erstellung des polnischen Teils des GeWiss-Korpus.

## 2. Die Aufbereitung des GeWiss-Korpus

Die theoretischen Annahmen des im Rahmen des GeWiss-Projektes entstehenden Korpus wurden schon von Fandrych et al. (2009) darge-

stellt. GeWiss[1] (Gesprochene Wissenschaftssprache kontrastiv) ist ein internationales Forschungsprojekt, welches das Ziel verfolgt, eine empirische Grundlage für eine vergleichende Untersuchung der gesprochenen Wissenschaftssprache des Deutschen, des Englischen und des Polnischen zu schaffen. Zu diesem Zweck wird ein Korpus erstellt, das zwei zentrale Genres der gesprochenen Wissenschaftssprache erfasst: Vortrag sowie Prüfungsgespräch. Aufgenommen werden dabei jeweils MuttersprachlerInnen und NichtmuttersprachlerInnen des Deutschen. Darüber hinaus werden Aufnahmen zum Englischen und Polnischen gemacht, vgl.:

Tabelle 1: Geplanter Korpusaufbau

| Sprache und Aufnahmeort | Vortrag | Prüfungsgespräch |
|---|---|---|
| Deutsch im deutschen Wissenschaftsbetrieb (40 h) | 10 h Experten (L1 Deutsch)<br>5 h Studenten (L1 Deutsch)<br>5 h Studenten (L2 Deutsch) | 10 h Studenten (L1 Deutsch)<br>10 h Studenten (L2 Deutsch) |
| Deutsch im englischen Wissenschaftsbetrieb (20 h) | 5 h Experten (L1 Englisch)<br>5 h Studenten (L1 Englisch) | 10 h Studenten (L1 Englisch) |
| Englisch im englischen Wissenschaftsbetrieb (20 h) | 5 h Experten (L1 Englisch)<br>5 h Studenten (L1 Englisch) | 10 h Studenten (L1 Englisch) |
| Deutsch im polnischen Wissenschaftsbetrieb (20 h) | 5 h Experten (L1 Polnisch)<br>5 h Studenten (L1 Polnisch) | 10 h Studenten (L1 Polnisch) |
| Polnisch im polnischen Wissenschaftsbetrieb (20 h) | 5 h Experten (L1 Polnisch)<br>5 h Studenten (L1 Polnisch) | 10 h Studenten (L1 Polnisch) |

Ein in dieser Weise entworfenes Korpus ermöglicht den Vergleich von Vorträgen/Referaten (die einerseits von Experten, also von Professoren, Doktoren und Doktoranden, andererseits von Bachelor- und Masterstudenten gehalten werden) und Prüfungsgesprächen. Der Themenkomplex der erhobenen Aufnahmen soll sich mit drei Feldern der klassischen Philologie decken, nämlich mit der Literaturwissenschaft, der Kulturwissenschaft und der Didaktik. Der Aufsatz handelt auch von den wichtigsten Parametern der Geräte, die bei der Aufnahme verwendet werden.

[1] Vgl. https://gewiss.uni-leipzig.de/de/.

## 3. Der polnische Teil des GeWiss-Korpus – Datenmanagement

Nach den oben genannten Annahmen umfasst der polnische Teil des GeWiss-Korpus 40 Aufnahmestunden, wobei 20 von ihnen deutschsprachige Aufnahmen sind, bei denen Polnisch L1 ist, die weiteren 20 Stunden machen polnischsprachige Aufnahmen aus. Die Aufbereitung eines so detaillierten und differenzierten Sprachmaterials erfordert einen großen Zeitaufwand. Damit ist ohne Zweifel das Problem des Datenmanagements verbunden, da die Aufnahmen allein lediglich der Ausgangspunkt sind: Bevor sie zum Korpus integriert werden, müssen sie eine Reihe von Bedingungen erfüllen.

Da die polnische Projektgruppe aus acht Personen besteht, die in unterschiedlichem Maße in den Projektarbeiten tätig sind, war es notwendig, einen Mechanismus zur Bestimmung des Vollständigkeitsgrades der durchgeführten Arbeiten zu schaffen. Das selbstständige Erheben und Zusammenstellen von Daten zu den durchgeführten Arbeiten durch die einzelnen Projektteilnehmer sowie die Anwendung der einer speziellen Schulung erfordernden Informationstechnik erweisen sich als unpraktisch. Als ein angebrachtes (d.h. einfaches und die Korpusbedingungen erfüllendes) Werkzeug stelle sich der entsprechend formatierte Kalkulationsbogen Google Docs heraus. Dank der Aufbewahrung in der Wolke erlaubt er allen Beteiligten einen gleichzeitigen Zugang zu den Metadaten sowie die Überprüfung des aktuellen Arbeitsstandes. Jede Aufnahme wird einer der neun Kategorien (Expertenvorträge, studentische Vorträge und Prüfungsgespräche in drei Subdisziplinen der Philologie) zugeordnet. Da die an die einzelnen Aufnahmen gestellten Anforderungen im Verlauf der Korpuserstellung beachtet werden (sollen), sind bei jeder einzelnen Aufnahme elf Einträge im Null-Eins-System erforderlich. Elf Einsen in der Tabelle bedeuten, dass sich die bestimmte Aufnahme zur Einbeziehung ins Korpus eignet. Separat werden die Aufnahmen platziert, die, aus welchem Grund auch immer, als unbrauchbar eingestuft werden, vgl. den Kalkulationsbogenausschnitt zu einer der Arbeitsphasen:

Tabelle 2: Kalkulationsbogenausschnitt

| EV ling | Zeit | trs | trsck | com | comck | msa | msv | lnk | ppt | pptj | hdt | zgd |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| EV_PL_020 | 00:27:28 | 1 | 1 | 1 | 1 | 1 | | | 1 | 1 | | 1 |
| EV_PL_021 | 00:23:48 | 1 | 1 | 1 | 1 | 1 | | | 0 | 0 | | 1 |
| EV_PL_027 | 00:11:51 | 1 | 1 | 1 | 1 | 1 | | | 1 | 1 | | 1 |
| EV_PL_028 | 00:19:51 | 1 | | 1 | 1 | 1 | | | 1 | 1 | | 1 |
| EV_PL_029 | 00:15:16 | 1 | 1 | 1 | 1 | 1 | | | 0 | 0 | | 1 |
| EV_PL_032 | 00:18:09 | 1 | 1 | 1 | 1 | 1 | | | 1 | 1 | | 1 |
| EV_PL_033 | 00:19:23 | 1 | 1 | 1 | 1 | 1 | | | 1 | 1 | | 1 |
| EV_PL_038 | 00:15:53 | 1 | 1 | 1 | 1 | 1 | | | 1 | 1 | | 1 |
| EV_PL_040 | 00:15:50 | | | 1 | 1 | | | | 0 | 0 | | 1 |
| EV_PL_049 | 00:17:43 | 1 | 1 | 1 | 1 | 1 | | | 0 | 0 | | 1 |
| EV_PL_050 | 00:19:44 | | | | | | | | 1 | 1 | | 1 |
| EV_PL_068 | 00:17:12 | 1 | 1 | 1 | 1 | 1 | | | 1 | 1 | | 1 |
| Summe | 03:42:08 | 10 | 9 | 11 | 11 | 10 | 0 | 0 | 9 | 9 | 0 | 14 |

Insgesamt wurden vorschriftsgemäß etwas mehr als zwanzig Stunden für das polnische Teilkorpus aufgenommen:

Tabelle 3: Die Aufnahmelänge in den einzelnen Teilkorpora

| | | |
|---|---|---|
| EV ling / EV kult / EV dyd | 5:21:47 h | 20:07:34 |
| SV ling / SV kult / SV dyd | 3:45:37 h | |
| PG ling / PG kult / PG dyd | 11:00:10 h | |

In Wirklichkeit war die Aufzeichnung einer doppelt so großen (genau 43:03:53 h) Zeitspanne nötig. Dies bedeutet, dass fast 23 Aufnahmestunden – trotz aller Mühe – zurückgewiesen werden mussten. Die Gründe, die dazu beigetragen haben, lassen sich im Grunde in zwei Gruppen aufteilen: in technische und theoretische Fragen.

## 4. Praktische Fragen

Bei den praktischen Fragen, die letztendlich Einfluss auf die endgültige Form des GeWiss-Korpus haben, müssen Aufnahmeorganisation und Technisches erwähnt werden. Bei der Aufnahmedurchführung war die Einholung datengeschützter Einwilligungserklärungen von den Prüflingen, den Prüfern und den Konferenzteilnehmern von großer Bedeutung sowie die Notwendigkeit, den tatsächlichen Verlauf der Kommunikationssituation zu bewahren. Zu den technischen Fragen zählten dagegen in erster Linie: nicht funktionelles (in plus) Aufnahmegerät, Verlust wesentlicher pragmatischer Informationen im Verlauf der Aufnahmemaskierung,

schließlich Schwierigkeiten beim Einholen der Handouts sowie unerwartete Ereignisse.

### 4.1. Organisatorische Fragen

#### 4.1.1. Einholung der Einwilligungserklärungen

Die Etappe der theoretischen Aufbereitung/Erstellung des GeWiss-Korpus ermöglicht die Bestimmung des endgültigen Zieles – einer für zukünftige Forschungsarbeiten optimalen Datenerhebungsmethode. In diesem Zusammenhang scheint die Einholung von Einwilligungserklärungen zu Beginn der Korpusvervollständigung schlüssig zu sein. Im Falle von Korpora geschriebener Texte, die urheberrechtlich geschützt sind, ist es besonders wichtig. Auch bei der Nutzung der Texte zu wissenschaftlichen Zwecken (die rechtlich nicht verboten ist) ist es ebenfalls empfehlenswert, die Autoren, Herausgeber, Verwalter der Internetseiten etc. über den Sachverhalt zu informieren. Im Falle der Erstellung eines Korpus der gesprochenen Sprache kann dieser Vorgang etwas komplizierter ausfallen. Die Korpuserstellung aus den bereits vorhandenen Aufnahmen unterscheidet sich hinsichtlich des Urheberrechts oder des Datenschutzes nur wenig von einer Textsammlung. Die Aufbereitung und Erstellung völlig neuer Fachaufnahmen bringt einige weitere Probleme mit sich und dies ist die Spezifik des GeWiss-Korpus. Beispielsweise stießen wir auf den Konferenzen nur vereinzelt auf Widerwillen beim Einholen der Einwilligungen. Einer der Gründe dafür mag darin bestehen, dass die Probanden in der Regel erfahrene Referenten waren (zahlreiche frühere Auftritte, Vorträge an den Universitäten etc.). Ein weiterer Faktor war sicherlich die Tatsache, dass Konferenzvorträge zur öffentlichen Rede zählen und somit an eine möglichst breite Anzahl von Empfängern gerichtet sind. Auch die vortragenden Studenten betrachteten – trotz geringerer Erfahrung – die Aufnahmegeräte im Saal als etwas Selbstverständliches und Normales. Absagen gab es nur selten und betrafen die oft in Form eines Witzes geäußerten Befürchtungen zu den Videoaufnahmen. Weitere Probleme beim Einholen der Einwilligungen tauchten eher im Kontext anderer Konferenzteilnehmer auf, also der Personen, die keine Vorträge selbst gehalten haben, aber vorhatten, sich an der Diskussion zu beteiligen. Gleichzeitig wünschten sie sich jedoch nicht, dass ihre Anwesenheit im Saal dokumentiert würde. Viel mehr Schwierigkeiten bereitete das Einholen der Einwilligungen bei den Prüfungsgesprächen (obwohl hier lediglich Audioaufnahmen durchgeführt wurden). Der erste Grund war wohl die Tatsache, dass das institutionsgebundene Prüfungsgespräch per se eine angespannte Situation darstellt.

Eine Rolle spielen hier auch psychologische Faktoren: ein häufiges (nicht unbedingt berechtigtes) Gefühl von Wissenslücken, einer unzureichenden Vorbereitung, aber auch Angst vor Misserfolg, insbesondere dann, wenn ein Dritter die Aussage im Nachhinein abhören und analysieren soll. Und je näher die Prüfung rückte, desto schwieriger wurde es, die Einwilligungen von den Studenten zu bekommen. Das Informieren über die Korpusannahmen – die Anonymität der Sprecher und das Maskieren aller Eigennamen (Name, Ort, Ereignis u.Ä.), die den Sprecher womöglich entlarven könnten, blieb ohne Erfolg. Aus unserer Erfahrung geht hervor, dass man eine größere Anzahl der Einwilligungserklärungen dann erhalten konnte, wenn man die Studenten schon zu Beginn des Kurses über das Projektvorhaben ausführlich informierte. Zu diesem Zeitpunkt lag die Prüfungstermin noch in weiter Ferne. Die Bedenken seitens der Studenten waren nicht der einzige Grund für die Schwierigkeiten bei der Vervollständigung der angestrebten Aufnahmen. Es war auch die Sorge um die Studenten selbst. Mit diesen Worten motivierten die Prüfer am häufigsten ihre Absage, die Prüfungsgespräche aufzuzeichnen. Mit den fortschreitenden Aufnahmen wuchsen auch die „Universitätslegenden" (in Anlehnung an die Stadtlegenden) über mutmaßliche und heimliche Zwecke dieser Aufnahmen, was die Sprechereinstellung den Aufnahmen gegenüber veränderte. Die Projektteilnehmer konnten diese Informationen lange Zeit nicht richtig stellen, da sie einfach nichts davon wussten.

### 4.1.2. Verfahrensprobleme bei der Durchführung der Aufnahmen

Die meisten unerwarteten Probleme bereitete, als die Aufnahme endlich zustande kam, der Versuch der Beibehaltung möglichst großer Natürlichkeit der gegebenen kommunikativen Situation. Beim Prüfungsgespräch durfte keiner der Projektteilnehmer im Saal anwesend sein. Im Endeffekt kam es öfter zu technischen Problemen (Wechsel des Audiogerätes, der Batterien bzw. der Speicherkarte). Diese Tätigkeiten an sich sind nichts Außergewöhnliches, man sollte hier jedoch die Spezifik des Prüfungsgesprächs im Bereich der Philologie am Standort Wrocław berücksichtigen. In der Praxis gab es – ab dem Zeitpunkt, als die erste Person den Raum betreten hat – keinen einzigen Moment einer natürlichen Pause in der Kontinuität des Prüfungsgesprächs, in der man das Audiogerät austauschen konnte, ohne in den Fragen-Antworten-Verlauf einzugreifen. Jede weitere Person, die den Raum betritt, bekommt vom Prüfer eine Anzahl von (oft mündlichen) Prüfungsfragen, auf die sie die Antworten vorbereitet, während die vorherige Person die Prüfung ablegt. Die Aufnahme eines ganzen Prüfungstages verursacht, dass eine spätere technische Bearbeitung der

Audiodateien notwendig ist. Dies ist wiederum auch von technischer Seite kein großes Problem – jedoch von der theoretischen Seite (eine Aufnahme – eine ganze Prüfung). Wo beginnt denn und wo endet jedes weitere Prüfungsgespräch? Das Erhalten der Fragen und deren Beantwortung werden oft durch die Aussage einer dritten Person unterbrochen. Auf diese Weise kam es oftmals zu Einschüben, Unterbrechungen und Eingriffen. Wenn dazu noch einer der Kandidaten kein Einverständnis für die Aufnahme erteilt, können viele Aufnahmeminuten nicht in das Korpus aufgenommen werden. Es ist nämlich nicht möglich, die Integrität der folgenden Aussage ohne ein künstliches Ausschneiden bedeutender Fragmente der realen Kommunikationssituation zu bewahren. Wie es scheint, ist die Hilfsbereitschaft seitens des Prüfers, der sich zum Ein- und Ausschalten des Aufnahmegerätes verpflichtet, auch keine gute Lösung. Die mit großer Mühe erhobenen Einwilligungserklärungen ziehen sehr oft keine Audiodatei nach sich, da die Geräte falsch bedient wurden. Als das am häufigsten auftretende organisatorische Problem bei den Konferenzaufnahmen (vor allem bei den Expertenvorträgen) erwies sich die Verletzung der Aussagekontinuität wegen technischer Probleme mit den Geräten oder mit deren Bedienung (z. B. die Kompatibilität der Software mit der vorbereiteten Präsentation o.Ä.), was oft zur Verwirrung führte: Sollte die zu lange dauernde Pause ausgeschnitten werden oder verursacht das einen zu großen Eingriff in die Aufnahme?

## 4.2. Technische Probleme

### 4.2.1. Unbrauchbarkeit der Aufnahmegeräte

Entscheidend für die Gestaltung eines effektiven Korpus ist die Wahl eines geeigneten Aufnahmegerätes. Es ist logisch und sehr wichtig, dass man bei jeder Korpuserstellung ein angemessenes Audio-/Videogerät zur Verfügung hat. Es sollte vor allem funktional sein. Die Funktionalität verstehen wir als eine Resultante der Qualität und der Einfachkeit bei der Bedingung. Es ist komplett unangebracht, nicht digitale Datenträger zu verwenden. Beispielsweise zieht die Verwendung einer nicht digitalen Kamera eine doppelt so lange Aufnahmedauer nach sich, da man die Bilder zuerst auf eine Kassette überspielen muss, über die Bearbeitung der gewonnenen Bilder ganz zu schweigen. Während der Aufnahme vervielfacht es dagegen die Anzahl der Personen, die mit der Aufnahme beschäftigt sind, weil es einen häufigen Wechsel der Kassette erfordert (üblicherweise nach jedem Vortrag/Referat). Bei ungünstigen Verhältnissen führt dies dazu, dass man nicht einmal einen Konferenzauftritt in voller

Länge aufnehmen kann. Dies verhindert wiederum *de facto* die Korpuserstellung, weil man unvollständige Vorträge/Referate in das Korpus nicht aufnehmen darf. Eine offene Frage ist auch die Qualität, in der das Gerät die Aufnahmen machen soll. Natürlich sollte sie so gut wie nur möglich sein. In der Praxis jedoch (wenn es sich um eine beträchtliche Größe des Korpus handelt) zieht eine hohe Aufnahmequalität die Vervielfachung des durch die Datenträger verwendeten Aufnahmeplatzes nach sich. Wenn die Aufnahme eine HD vergleichbare Qualität haben soll, verlangt sie für die spätere Bildbearbeitung bessere (sprich teurere) Prozessoren und Grafikkarten. Darüber hinaus beinhaltet eine qualitativ hohe (Audio-)Aufnahme verschiedene Arten von akustischen Phänomenen, die die Aufmerksamkeit der Zuhörer in keinster Weise stört, wobei der/die Transkribierende bei der Verschriftlichung zurecht schlussfolgert, dass sie einen erheblichen Einfluss auf die Kommunikationssituation haben konnten. Die spätere Transkription der erhobenen Aufnahmen ist dadurch reich an Annotationen. Es ist daher notwendig, ein sinnvolles Gleichgewicht zwischen der Genauigkeit der Audioaufnahmen und der durch das Mikrofon aufgefangenen akustischen Phänomene zu finden.

### 4.2.2. Probleme bei der Bearbeitung von Vorträgen

Die Aufnahmen sollten – in Anlehnung an die theoretischen Annahmen des Projekts – mit allerlei Medien versehen werden, die der/die Vortragende/r benutzt (multimediale Präsentation, Audiodateien, Handouts). Mehrmals wollten die Vortragenden ihre Zusatzmaterialien nicht vor Ort übergeben, sondern diese nach der Konferenz per E-Mail schicken. Trotz zahlreicher Anfragen und Bitten ist eine Anzahl der Vortragenden diesem Versprechen nicht nachgekommen.

### 4.2.3. Probleme bei der Verarbeitung von Aufnahmen

Eine der wichtigsten Anforderungen bei der Korpus-Erstellung ist die Anonymität. Die Mehrheit der Aufnahmen beinhaltet Phrasen, die maskiert werden müssen, sonst besteht die Möglichkeit, dass die Sprecher erkannt werden. Das Maskieren beruht auf der Änderung des Aufnahmefragmentes in ein Rauschen, das das Erkennen der Original-Aussage unmöglich macht. Bei der Transkription wird so ein Fragment durch ein ähnlich klingendes Fragment ausgetauscht: z.B. durch einen Vor- und Nachnamen mit genau derselben Silbenanzahl und mit denselben Anfangsbuchstaben (vgl. Fandrych et al. 2012). Oft trägt jedoch ein z.B. von dem Prüfer ausgespro-

chener Vor- oder Nachname eine wesentliche Zusatzinformation mit sich. Nach dem Ersetzen dieses Fragmentes durch ein Rauschen gehen alle intonatorischen und prosodischen Informationen verloren, die auf die emotionelle Verfassung des Sprechers verweisen und eine gewisse Interpretation des weiteren Gesprächsverlaufs erzwingen. Der/die Transkribierende/r, der/die eine solch maskierte Aufnahme bekommt, hat keine Ahnung davon, ob der Prüflingsname mit Verärgerung, Empörung oder etwa mit Nachsicht ausgesprochen wurde. Das scheint vielleicht unwesentlich zu sein, aber in Wirklichkeit macht es eine solide Untersuchung der Prüfer-Prüfling-Beziehung unmöglich. Eine Lösung kann hier eine Ergänzung der maskierten Fragmente um eine zusätzliche Information sein, jedoch wäre sie immer das Ergebnis einer individuellen Interpretation, unüberprüfbar durch einen breiteren Kreis von Forschern. Die Maskierung von Konferenzvorträgen bleibt ebenfalls umstritten. Es ist alles andere als schwierig, die nötigen Informationen über die Sprecher zu erwerben, den Namen an das Thema anzupassen oder den Sprecher auf einer Videoaufnahme zu erkennen. Eine zusätzliche Maskierung der Konferenzpräsentationen beeinträchtigt oftmals das grafische Konzept, weil einige Präsentationselemente (die beispielsweise von der Hochschule des Sprechers verlangt werden und ein System der visuellen Identifikation schaffen) entfernt werden müssen. Dies beeinflusst in gewisser Weise die Verteilung der Elemente auf einem Dia und das Kolorit des Ganzen. Dazu kommen die chromatischen Strategien bei der Konferenzkommunikation, die ohne Kontext ihre Bedeutung verlieren: Warum diese und keine andere Farbe, warum diese und keine andere Konstellation?

### 4.2.4. Unerwartete Ereignisse

Eine inhärente Eigenschaft bei der Korpuserstellung sind unerwartete Ereignisse. Es ist natürlich unmöglich, sie vollkommen zu eliminieren, sie sollten jedoch bei der zeitlichen Einschätzung der Datenaufbereitung berücksichtigt werden. Sie können nämlich den Zeitpunkt der Korpusveröffentlichung wesentlich verzögern. Dazu gehören z. B. unabsichtliche Zusammenstöße bzw. Verschiebungen der Aufnahmegeräte oder der Kabel, wie auch eine zufällige (manchmal nur zeitweilige) Anwesenheit im Bild der Kamera. Es sind belanglose Ereignisse, die jedoch ab und an darin resultieren, dass die Aufnahme zurückgewiesen wird. Dies macht (je nach Dauer) ca. 6,5% des gegebenen Korpusteils aus.

## 5. Theoretische Fragen

Der zweite Fragenkomplex, der in entscheidender Weise die Korpusform beeinflusst hat, ist rein theoretischen Ursprungs. Es ist jedoch unentbehrlich, diesen Komplex während der Korpusaufbereitung mitzuberücksichtigen. Es geht hier vor allem um Angelegenheiten, die mit der Kategorisierung des Sprachmaterials verbunden sind. Diese ergeben sich hauptsächlich aus interkulturellen und institutionell-organisatorischen Unterschieden. Mit anderen Worten: Es geht um Angelegenheiten, die – wenn überhaupt – bei der Erstellung eines einsprachigen (einkulturellen) Korpus vorkommen.

### 5.1. Kategorisierung philologischer Unterdisziplinen

Die erste Frage, die ein Problem bei der Erstellung eines vergleichenden Korpus entstehen lässt, ist die angenommene Kategorisierung der Unterdisziplinen. Am Beispiel der Polonistik in Wrocław entspricht diese Aufteilung nur teilweise der wirklichen Beteiligung des gegebenen Elements im philologischen Studium. Sichtbar wird dies etwa am Beispiel der institutionellen Organisation des Instituts, wo auf 13 literatur- und sprachwissenschaftliche Einrichtungen (etwa 80 Mitarbeiter) nur eine Didaktik-Einrichtung (3 Mitarbeiter) kommt. Dies führt zu einer viel niedrigeren Anzahl von adäquaten Prüfungen am Institut. Es hat auch einen gewissen Einfluss auf die Aufzeichnung von studentischen Vorträgen. Dieses Problem verlangt letzten Endes drei Entscheidungen in Bezug auf das erhobene Sprachmaterial:

- eine vollkommene Übereinstimmung mit dem tatsächlichen Sachverhalt der polnischen Philologie auf Kosten der kleineren Vergleichbarkeit von mehrsprachigen Korpora,
- eine vollkommene Vergleichbarkeit mit anderen Korpora auf Kosten der Adäquatheit hinsichtlich einer realen Beteiligung dieser Art von Texten im polnischen Wissenschaftsbetrieb,
- der Versuch eines Gleichgewichtes zwischen den oberen.

Die erste Entscheidung erscheint im Fall eines vergleichenden Korpus absolut ziellos. Die zweite Entscheidung fällt ebenfalls aus: Das Hauptziel des Projektes ist die Beschreibung der deutschen Wissenschaftssprache „in Beziehung zu". Bei der Erstellung des polnischsprachigen Teilkorpus entschieden wir uns für die dritte Variante, nämlich die Berücksichtigung einer – zwar etwas überhöhten, aber kleineren als angenommen – Sammlung von Texten, die auf Didaktik hin orientiert sind.

### 5.2. Kategorisierung des Promotionsstudiums

Als ein weiteres unerwartetes Problem erwies sich die Kategorisierung des Promotionsstudiums. Nach den Korpusannahmen werden Vorträge von Doktoranden als Expertenvorträge aufgefasst. Allerdings ist der Doktorandenstatus im polnischen Rechts- und Bildungssystem nicht eindeutig; die Bilanz von Gewinn und Verlust vonseiten der Institute bevorzugt üblicherweise die Einstufung der Doktoranden als Studenten. Wenn man diesen Zustand auf die wichtigsten Fragen des Korpus überträgt, sollte man bemerken, dass wir *de facto* mit zwei Arten von Vorträgen (der Doktoranden) zu tun haben – mit denen, die entweder auf Expertenkonferenzen oder auf Studentenkonferenzen gehalten wurden. Die Kategorisierung erster Art (als Expertenvorträge) weckt keine Vorbehalte. Bleibt die Frage nach der Kategorisierung zweiter Art im Hinblick auf die vorgetragenen Texte. Bei der Erstellung des polnischsprachigen Teilkorpus bemühten wir uns, wo immer es nur möglich war, aus dem Expertenquantum der Vorträge diejenigen auszuschließen, die von einer Studentenkonferenz kamen. Diese Entscheidung ist jedoch in vieler Hinsicht arbiträr.

### 5.3. Kategorisierung der Kommunikationsereignisse und/oder -arten

Die wichtigste Frage im Hinblick auf den Aufbau eines vergleichenden mehrsprachigen Korpus ist die Frage nach der Kategorisierung der Genres. Das Phänomen ihrer gewissen interkulturellen Inkongruenz ist allgemein bekannt (vgl. z.B. Wierzbicka 1985) und wurde im Bereich der Wissenschaftskommunikation behandelt (vgl. z.B. Duszak 1997, für das Polnische z.B. Duszak 1998). Bei der Erhebung des Sprachmaterials erwies sich die Frage, sowohl bei den monologischen wie auch dialogischen Texten, als problematisch. Nach dem Korpusdesign ist die anschließende Diskussion ein untrennbarer Teil eines jeden gehaltenen Vortrags. Im Grunde stehen diese zwei Genres in einer engen Verbindung zueinander. Man kann sie *de facto* als Unterkomponenten eines größeren Ganzen betrachten, also als Konferenzauftritt. Demzufolge kann man sie, wenn sie nur nacheinander auftreten, in Form eines einzigen – deutlich reicheren als der Vortrag selbst – Kommunikationsereignisses registrieren. So eine Einstellung wird erst dann kompliziert, wenn man auf eine andere Organisation des Makrogenres trifft, nämlich auf die Konferenz. Drei prototypische Modelle illustriert die Tabelle unten:

Tabelle 4: Die prototypischen Organisationsmodelle eines Konferenzverlaufes

| Vortragsmodell | | Gemischtes Modell | | Blockmodell | |
|---|---|---|---|---|---|
| | | Block 1 | | Block 1 | |
| Vortrag 1 | Konferenz-vortrag | Vortrag 1.1 | Konferenz-vortrag | Vortrag 1.1 | Konferenz-vortrag |
| Diskussion 1 | | Diskussion 1.1 | | Vortrag 1.2 | |
| Vortrag 2 | Konferenz-vortrag | Vortrag 1.2 | Konferenz-vortrag | Vortrag 1.3 | |
| Diskussion 2 | | Diskussion 1.2 | | Diskussion 1 | |
| Vortrag 3 | | Block 2 | | Block 2 | |
| (…) | | (…) | | (…) | |

Der unterscheidende Punkt ist hier die Abwesenheit eines Themenbereiches (der sich mit dem rein organisatorischen Bereich zwar decken kann, aber nicht muss). Das erste der dargestellten Modelle, hier als *Vortragsmodell* bezeichnet, hat einen außerordentlich individualistischen Charakter – er fasst die Konferenz als eine Sammlung von aufeinander folgenden, jedoch deutlich voneinander abgetrennten Konferenzauftritten, zusammen. Diese Auftritte können selbstverständlich miteinander verbunden sein, müssen es aber nicht. Der Gegensatz zu diesem Modell ist das kollektive *Blockmodell*, bei dem eine Reihe von thematisch miteinander verbundenen Vorträgen als eine Ganzheit angesehen wird. Angesichts dieser Ganzheit entsteht erst eine Diskussion. Das Zwischenmodell, als *gemischt* bezeichnet, setzt die Existenz von deutlichen Themenbereichen voraus, jedoch mit Einhaltung einer individualistischen Einstellung zum Vortrag.

Wenn es sowohl im Projekt geplanten Korpusmodell (Vortrag) als auch im gemischten Modell keine Schwierigkeiten gibt, den ganzen Konferenzauftritt aufzunehmen (Vortrag + Diskussion), so entstehen Probleme im Blockmodell, dem populärsten Modell, wenn es um polonistische Konferenzen geht. Jede der möglichen Auswege ist unbefriedigend:

- die Eintragung des gesamten Blocks mit der anschließenden Diskussion ist adäquat hinsichtlich eines echten Kommunikationsereignisses, es verhindert jedoch den Vergleich einzelner Genres unter dem Blick einfachster Parameter (z. B. Diskussionsdauer);
- die Eintragung einzelner Vorträge, darunter den letzten mit anschließender Diskussion, ist zum Teil inkohärent (in zwei Aufnahmen fehlt die Diskussion, in der dritten Aufnahme bezieht sie sich auf Vorträge, die in einer anderen Aufnahme registriert wurden);
- die Eintragung einzelner Vorträge samt einer künstlichen Hinzufügung entsprechender Fragmente der Diskussion zerstört vollkommen die Datenauthentizität. Die Lösung scheinen der Verzicht auf

die Eintragung der Diskussion und die Einschränkung dieses Korpusfragments nur auf die Vorträge zu sein.

### 5.4. Kategorisierung der gesprochenen Sprache (Grad der Mündlichkeit)

Eine genauso problematische Frage ist im Falle von wissenschaftlichen Vorträgen die Kategorisierung des Begriffes der sog. gesprochenen Sprache. Die offensichtlichste Lösung ist die im Projekt angenommene Definition des *gesprochenen* als *gesagten*. Auf diese Weise kann man ein gewisses Mündlichkeitskontinuum festlegen. Es verläuft von dem (beinahe) spontanen Auftritt, durch eine vorbereitete (protoptypische) und auswendig reproduzierte (klassische rhetorische) Rede bis hin zu einem Vortrag, der vollständig abgelesen wird. Solch eine Auffassung gewährleistet eine maximale Authentizität der im Korpus gesammelten Daten, weil dadurch keine Konferenzvorträge abgelehnt werden, die man als die „mit den Annahmen widersprüchlich" bezeichnet. Gleichzeitig verursacht die Überzeugung jedoch weitere Komplikationen bei der automatischen Korpusdatenverarbeitung. Wenn wir nämlich annehmen, dass wir im Korpus gesprochene Textbeispiele haben und das Korpus weitgehend aus den vorgelesenen schriftlichen Texten bestehen wird (was bei den polonistischen Konferenzen der Fall ist), bekommen wir im Grunde ein verfälschtes Bild „dessen, was gesagt wird", und zwar auf beliebiger Sprachebene. Mit anderen Worten: Das Sprachmaterial wird unweigerlich die Merkmale der geschriebenen, nicht der gesprochenen Sprache besitzen.

## 6. Schlussfolgerungen

Die oben dargestellten Fragen kann man in fünf synthetischen Schlussfolgerungen zusammenfassen:

1. Selbst bei dem genausten Korpusentwurf wird seine endgültige Gestalt von einer Reihe unvorhersehbarer Faktoren abhängen.
2. Der Grad der Vorhersehbarkeit wächst sowohl im Fall der Erstellung eines Korpus der gesprochenen Sprache (praktische Fragen) wie auch eines vergleichenden (theoretisch-interkulturelle Fragen).
3. Zu den wesentlichen Faktoren von praktischer Seite her, die die endgültige Form des Korpus beeinflussen, zählen vor allem organisatorische Fragen (die Aufnahmedurchführung) und eine Reihe rein technischer Fragen, die im Laufe der Korpusarbeiten entstehen.

4. Unter den wesentlichen Faktoren von theoretischer Seite her sollte man die Kategorisierung zweier Arten von sprachlichen Phänomenen erwähnen: die außertextuellen, die die Korpusform beeinflussen und die strikt intertextuellen. Im ersten Fall geht es um den Status einzelner philologischer Unterdisziplinen und des Promotionsstudiums in einzelnen akademischen Kulturen. Im zweiten Fall um die Kategorisierung scheinbar parallelen Geres (z.B. Konferenzauftritt / Vortrag) und ihre fundamentalen Eigenschaften (z.B. Grad der Mündlichkeit).
5. Selbst bei dem genausten Design eines so komplizierten Korpus, wie es das vergleichende Korpus der gesprochenen Sprache ist, und bei Berücksichtigung aller empirischen Daten, muss ein Großteil von Entscheidungen über die endgültige Korpusform einen arbiträren Charakter annehmen.

## Literatur

Duszak Anna (Hrsg.), 1997, Culture and styles of academic discourse, Berlin.

Duszak Anna, 1998, Tekst, dyskurs, komunikacja międzykulturowa, Warszawa.

Fandrych Christian / Tschirner Erwin / Meissner Cordula / Rahn Stefan / Slavcheva Adriana, 2009, Gesprochene Wissenschaftssprache kontrastiv: Deutsch im Vergleich zum Englischen und Polnischen. Vorstellung eines gemeinsamen Forschungsvorhabens, in: Studia Linguistica 28, Wrocław, S. 7–30.

Fandrych Christian / Meissner Cordula / Slavcheva Adriana, 2012, The GeWiss Corpus: Comparing Spoken Academic German, English and Polish, in: Schmidt T./Wörner K. (Hrsg.), Multilingual corpora and multilingual corpus analysis, Amsterdam, S. 319-337.

Köhler Reinhard, 2005, Korpuslinguistik. Zu wissenschaftstheoretischen Grundlagen und methodologischen Perspektiven, in: LDV-Forum 20/2, S. 1-16.

Wierzbicka Anna, 1985, Different cultures, different languages, different speech acts: Polish vs. English, Journal of Pragmatics 9 (2-3), S. 145-178.