Referenzkorpus der deutschen Sprache, sondern um ein Referenzkorpus der deutschen Schriftsprache.

Das britische Nationalkorpus (BNC) hingegen, das das britische Englisch repräsentieren soll, beinhaltet zu 90% Texte der geschriebenen Sprache und zu 10% transkribierte gesprochene Sprache. Die Texte der Schriftsprache umfassen unter anderem Auszüge aus regionalen und überregionalen Zeitungen, Fachzeitschriften und Zeitschriften für alle Alters- und Interessensgruppen, wissenschaftliche und nichtwissenschaftliche Bücher, veröffentlichte wie unveröffentlichte Briefe und Notizen, Aufsätze von Schülern und Studenten. Bei den Texten der gesprochenen Sprache handelt es sich überwiegend um informelle spontane Gespräche, die von Freiwilligen aufgenommen wurden, daneben aber auch um gesprochene Sprache aus anderen Kontexten wie Regierungsgespräche oder Radiosendungen. Wenn man sich allerdings klar macht, welchen Anteil an der tagtäglichen Kommunikation die gesprochene Sprache einnimmt, erscheint die gesprochene Sprache auch mit einem Anteil von 10% noch deutlich unterrepräsentiert zu sein. Für eine Untersuchung zur gesprochenen Sprache reichen die im BNC enthaltenen 10 Millionen Textwörter dennoch gut aus.

Spezialkorpora erheben hingegen nicht den Anspruch, repräsentativ für eine Sprache in ihrer Gesamtheit zu sein. Sie dienen vielmehr dazu, eine bestimmte Varietät der Sprache wie die Jugendsprache, die deutsche Rechtssprache, die Zeitungssprache, das Hessische oder die Sprache von Deutsch-als-Fremdsprache-Lernern zu erforschen. Diesen Teilbereich der Sprache sollten sie jedoch hinreichend repräsentieren.

Viele bekannte Korpora sind Spezialkorpora, die auf bestimmten Textsorten basieren. Insbesondere Korpora der Zeitungssprache sind weit verbreitet. Eine große Zahl der IDS-Korpora wie das Mannheimer-Morgen-Korpus oder das Bonner Zeitungskorpus bestehen aus Zeitungstexten. Auch verschiedene Baumbanken wie das TIGER-Korpus und das NEGRA-Korpus sind Korpora der Zeitungssprache. Daneben gibt es Korpora, die auf andere Textsorten spezialisiert sind, etwa auf Lyrik, Romane, Gebrauchsanweisungen oder Bibelübersetzungen.

Als Spezialkorpora einzuordnen sind aber auch die bereits erwähnten Lerner- und Spracherwerbskorpora. Lernerkorpora wie das Fehler-annotierte Lernerkorpus des Deutschen als Fremdsprache (FALKO), das derzeit in Berlin aufgebaut wird, enthalten Texte, die von Schülern und Studenten in einer Fremdsprache verfasst wurden (vgl. Kapitel 1.5). Spracherwerbskorpora wie die CHILDES-Korpora und das Saarbrücker Korpus der Kindersprache umfassen Transkripte und Aufzeichnungen von – in der Regel gesprochener – Kindersprache (vgl. Kapitel 2.7).

Eine große Zahl an Spezialkorpora sind Fachtextkorpora wie das Darmstädter Korpus deutscher Fachsprachen, das rund 2,8 Millionen Textwörter aus den Gebieten Bauingenieurwesen, Elektrotechnik, Maschinenbau und Wirtschaft enthält. Weitere Fachtextkorpora sind die bereits erwähnten Korpora zum Computerdiskurs und zur Sprache der Bochumer Stadtverwaltung (vgl. Kapitel 1.5). Ein mehrsprachiges Fachtextkorpus ist das OPUS-Korpus, das Teilkorpora zur Verwaltungssprache und zu verschiedenen technischen Disziplinen umfasst (vgl. Kapitel 2.10). Als Beispiele für historische Fachtextkorpora sei an dieser Stelle lediglich auf zwei frühneuhochdeutsche Korpora, nämlich das Olmützer medizinische Korpus sowie das Erlanger Dürer-Korpus mit mathematisch-technischen Texten, verwiesen.

2.10 Einsprachige und mehrsprachige Korpora

Die meisten Korpora enthalten nur Daten aus einer Sprache. Dies gilt auch für Referenzkorpora wie DEREKO oder BNC, die zwar Material zu einer Vielzahl von Varietäten beinhalten, die jedoch alle derselben Sprache, in diesem Fall dem Deutschen bzw. Englischen, zuzuordnen sind (vgl. Kapitel 2.9). Daneben gibt es aber auch Korpora wie das International Sample of English Contrastive Texts (INTERSECT) oder das Chemnitzer German/English-Translation-Korpus, die sowohl deutsche als auch englische Texte enthalten. Das Verbmobil-Korpus umfasst neben deutschen und englischen Texten auch japanische Texte. Eine besonders große sprachliche Vielfalt bietet das OPUS-Korpus. Die Anzahl der enthaltenen Sprachen schwankt in den fünf Teilkorpora zwischen sechs (OpenOffice-Korpus) und 61 (KDE-Korpus). Eine Vielzahl mehrsprachiger, allerdings kostenpflichtiger Korpora bietet die Evaluations and Language Resources Distribution Agency (ELDA).

Bei **mehrsprachigen Korpora** ist zu unterscheiden zwischen so genannten Parallelkorpora und vergleichbaren Korpora. **Parallelkorpora** wie das Chemnitzer German/English-Translation-Korpus zeichnen sich dadurch aus, dass sie Originaltexte in einer Sprache und deren Übersetzung in eine oder mehrere andere Sprachen beinhalten. Das Chemnitzer Korpus umfasst insgesamt rund zwei Millionen Textwörter, je eine Million Textwörter für das Deutsche

und das Englische, die aus den Bereichen Politik, Wissenschaft und Tourismus stammen. Zugänglich ist das Korpus über die Internetseite der Chemnitzer Internet-Grammatik.

Wie das Chemnitzer German/English-Translation-Korpus ist auch das OPUS-Korpus ein Parallelkorpus, das im Internet frei zugänglich ist. Es enthält Gebrauchsanweisungen und Dokumente der Europäischen Union mit den jeweiligen Übersetzungen. Dabei kann ausgewählt werden, in wie viele Sprachen eine Textpassage übersetzt werden soll. In Abbildung 2 werden die deutsche, englische, französische und finnische Version eines Satzes angezeigt.

11014334 Als Katalane würde ich mir für die Zukunft wünschen, dass auch meine **Sprache**, die von zehn Millionen europäischen Bürgern gesprochen wird, hier im Hause offiziell anerkannt wird.

en As a native Catalan, it is my hope that, in the future, my language, which is spoken by 10 million European citizens, may also be recognised in this Chamber.

fi Syntyperäisenä katalonialaisena toivon, että tulevaisuudessa äidinkieleni, jota puhuu 10 miljoonaa Euroopan kansalaista, tunnustetaan myös tässä parlamentissa.

fr En tant que Catalan, je souhaiterais que, demain, ma langue, qui est celle de dix millions de citoyens européens, ait également droit de cité dans cette maison.

Abbildung 2: OPUS-Korpus: Konkordanz für *Sprache* (Ausschnitt)

Charakteristisch für **vergleichbare Korpora** ist, dass alle Teilkorpora denselben Aufbauprinzipien folgen. Die Teilkorpora verfügen also über eine identische Struktur. Im Gegensatz zu Parallelkorpora, die prinzipiell mehrsprachig sind, können vergleichbare Korpora sowohl ein- als auch mehrsprachig sein.

Ein Beispiel für ein mehrsprachiges vergleichbares Korpus ist das PAROLE-Korpus. Es besteht aus zwölf Teilkorpora, unter anderem einem deutschen, englischen und französischen Teilkorpus, mit jeweils rund 20 Millionen Textwörtern. Drei weitere Teilkorpora enthalten eine geringere Anzahl an Textwörtern. Alle Teilkorpora wurden nach einheitlichen Kriterien aufgebaut und mit zusätzlichen grammatischen Informationen versehen. In jedem Teilkorpus wurden dieselben Anteile an Texten aus Büchern, Zeitungen und Zeitschriften erhoben. Ein Beispiel für ein mehrsprachiges vergleichbares Korpus der gesprochenen Sprache ist das Verbmobil-Korpus, das deutsche, englische und japanische Spontansprache aus dem Bereich der Terminvereinbarung enthält.

Als vergleichbare einsprachige Korpora sind das Brown-Korpus, das Lancaster-Oslo/Bergen-Korpus (LOB) und das Kolhapur-Korpus zu nennen. Alle drei Korpora verfügen über dieselbe Struktur, nämlich die des Brown-Korpus, bilden jedoch unterschiedliche regionale Varianten des Englischen – amerikanisches, britisches und indisches Englisch – ab.

Aufgabe 9: Das International Corpus of English (ICE) enthält insgesamt 20 Millionen Textwörter. Das Korpus besteht aus zwanzig Teilkorpora aus Ländern, in denen Englisch die einzige oder eine der offiziellen Nationalsprachen ist. Jedes der Teilkorpora enthält zu 60% gesprochene und zu 40% geschriebene Sprache, die nach denselben Kriterien erhoben wurden.
Handelt es sich beim ICE Ihrer Meinung nach um ein einsprachiges oder ein mehrsprachiges Korpus? Handelt es sich bei den Teilkorpora um parallele oder vergleichbare Korpora? Bitte begründen Sie Ihre Ansicht.

2.11 Zusammenfassung

Korpora werden nach formalen Kriterien eingeteilt in computerlesbare und Papierkorpora, in Gesamt- und Teilkorpora, in Proben- und Volltextkorpora, in statische und Monitorkorpora sowie in annotierte und nicht annotierte Korpora.
Im Hinblick auf ihren Inhalt werden Korpora der gesprochenen und der geschriebenen Sprache, Korpora der Gegenwartssprache und historische Korpora, Referenz- und Spezialkorpora sowie ein- und mehrsprachige Korpora unterschieden.

Grundbegriffe: Annotation, annotiertes Korpus, Baumbank, computerlesbares Korpus, einsprachiges Korpus, historisches Korpus, Korpus der Gegenwartssprache, Korpus der geschriebenen Sprache, Korpus der gesprochenen Sprache, mehrsprachiges Korpus, Monitorkorpus, Papierkorpus, Parallelkorpus, Probenkorpus, Referenzkorpus, Spezialkorpus, statisches Korpus, Teilkorpus, vergleichbares Korpus, Volltextkorpus

Weiterführende Literatur
Sinclair (1998) gibt einen Überblick über verschiedene Arten von Korpora. Eine weitere Korpustypologie sowie eine systematische Übersicht über deutschsprachige Korpora bieten Lemnitzer/Zinsmeister (2006, Kapitel 5). Hinweise auf weitere Korpora des Deutschen, insbesondere historische und mehrsprachige Korpora, finden sich in dem Sammelband von Schwitalla/Wegstein (Hgg.) (2005).