DE GRUYTER OPEN

Acta Univ. Sapientiae, Philologica, **9**, **3,** (2017) 135–144

DOI: 10.1515/ausp-2017-0034

Zu theoretischen und praktischen Aspekten des Fachübersetzens

Sprachkorpora im Dienste der kulturellen Vielfalt

Olívia SEIDL-PÉCH
Institut für Fremdsprachen
Technische und Wirtschaftswissenschaftliche Universität Budapest (BME)
olivia@inyk.bme.hu

Abstrakt. In den letzten Jahrzehnten wird immer häufiger über die Korpuslinguistik geschrieben, die ihren Aufschwung vor allem dem Gebrauch elektronischer Korpora seit den 1960er-Jahren verdankt (Brown Korpus). Inzwischen benutzen immer mehr Teilbereiche der allgemeinen und angewandten Sprachwissenschaft (z. B. Computerlinguistik, Diskursanalyse, synchrone und diachrone Sprachwissenschaft, kontrastive Linguistik, Lexikologie und Lexikographie, Psycholinguistik, Soziolinguistik, die Sprachlern- und Lehrforschung, Übersetzungswissenschaft) korpuslinguistische Methoden. In der Sprachforschung wird vorwiegend der empirische und deskriptive Charakter der korpusgestützten Sprachanalyse betont.
Die Erstellung und Vorhaltung digitaler Sprachkorpora wurde dank der digitalen Revolution des 20. und 21. Jahrhunderts auch für kleinere Nationen und Sprachgemeinschaften, ebenso wie für Wissenschaftler/innen zugänglich. So können heutzutage Sprachkorpora nicht nur als Hilfsmittel der Sprachforschung und der Übersetzung(swissenschaft) betrachtet werden, sondern sie liefern auch einen Beitrag zur Bereicherung der kulturellen Vielfalt. Im Fokus des Artikels stehen neben internationalen Beispielen die wichtigsten ungarischen Korpora und die Kriterien der Korpuserstellung im Zusammenhang mit der kulturwissenschaftlichen Orientierung der Korpuslinguistik.

Schlüsselwörter: (Fach)Übersetzen, korpusgestützte Sprachanalyse, kulturelle Vielfalt, Textkorpora, Übersetzungsforschung

Abstract. In the past few decades, it has extensively been written about corpus linguistics, which has owned its upswing mainly to the use of electronic corpora since the 1960s (Brown Corpus). Meanwhile, an increasing number of fields within general and applied linguistics (e.g. computational linguistics, discourse analysis, contrastive linguistics, diachronic and

synchronic linguistics, language teaching and learning research, lexicology and lexicography, psycholinguistics, sociolinguistics, translation studies) have been using corpus linguistic methods. In linguistic research, the empirical and descriptive character of corpus-based linguistic analysis has also been given an emphasis.
Thanks to the digital revolution of the 20th and 21st centuries the creation and provision of digital linguistic corpora is becoming accessible for smaller nations and language communities as well as for scientists. Nowadays, linguistic corpora cannot only be regarded as a tool to support language research and Translation Studies, but they also contribute to the enrichment of cultural diversity. The article focuses on international examples as well as on the most significant Hungarian corpora. The paper also discusses the criteria of corpus creation and several cultural aspects of corpus linguistics.

Keywords: corpus-based language analysis, cultural diversity, technical translation, text corpora, translation research

1. Korpusevidenz

Die Übersetzungswissenschaft greift bezüglich der Korpuslinguistik (corpus-based translation studies) auf Vorläufer wie John Sinclair (1991) und Mona Baker (1993) zurück. Inzwischen hat sich die Disziplin entwickelt, ihre Zielsetzungen und Eigenschaften ausführlich definiert und ausgerichtet. Der Begriff Korpuslinguistik bezeichnet „die Beschreibung von Äußerungen natürlicher Sprachen, ihrer Elemente und Strukturen, und die darauf aufbauende Theoriebildung auf der Grundlage von Analysen authentischer Texte, die in Korpora zusammengefasst sind“ (Lemnitzer–Zinsmeier 2006. 9).

Wesentlichstes Merkmal des korpuslinguistischen Herangehens ist die Auswahl geeigneter Korpora, die bei der empirischen Untersuchung von Sprachen und Subsprachen und bei der Beschreibung sprachlichen und subsprachlichen Eigenschaften und Regularitäten den Ziel- und Forschungsfragen entsprechen. Ein Korpus ist eine „Sammlung einer möglichst hohen, notwendigerweise aber immer begrenzten Anzahl möglichst zusammenhängender sprachlicher Äußerungen (gesprochen oder/und geschrieben) aus möglichst natürlichen Kommunikationssituationen“ (Gluück 2005. 357).

Heutzutage sind Korpora auf Rechnern gespeichert und maschinenlesbar, diese beiden Faktoren ermöglichen eine genaue und schnelle Abfragung der Daten. Korpora werden nach vorab definierten Kriterien zusammengestellt, so dass sie ein genaues Abbild der untersuchten Sprache oder Subsprache darstellen. Bei der Datensammlung müssen daher die Kriterien der Repräsentativität und der Ausgewogenheit beachtet werden, wobei die Sammlung authentischen Sprachmaterials Voraussetzung für Forschungskorpora ist. Bei der Zusammensetzung

von Sprachkorpora können verschiedene Auswahlkriterien im Vordergrund stehen. Je nach Zielsetzung der empirischen Sprachbeschreibung oder der Anwendung bzw. des Auswertungsziels entstehen verschiedenen Korpustypen. Diese Auswahlkriterien bestimmen gleichzeitig die kulturwissenschaftliche Orientierung des Korpus.

2. Bedarf der maschinellen (Fach)Übersetzung

Das erste elektronische Korpus ist das Brown Corpus (Brown University Standard Sample of Present-Day American English) und wurde im 1946 von W. N. Francis und H. Kučera erstellt (Francis–Kučera 1979). Anschließend erschienen zahlreiche monolinguale Korpora (Albanian National Corpus, British National Corpus, Bulgarian National Corpus, Czech National Corpus, Corpus del Español, Hungarian National Corpus, Lancaster Corpus of Mandarin Chinese, National Corpus of Polish, Quranic Arabic Corpus, Romanian Balanced Corpus, Russian National Corpus, Slovak National Corpus, Szeged Treebank, TIGER Corpus, Turkish National Corpus), denen weitere bi- und multilingualen Korpora (Aligned Hansards of the 36th Parliament of Canada, COMPARA, English-Norwegian Parallel Corpus, European Parliament Proceedings 1996-2001) folgten. Zweifelsfrei ist die relativ rasche Verbreitung der Korpuslinguistik der maschinellen Übersetzung (machine translation) zu verdanken. Die automatische Übersetzung wurde durch korpusbasierte Methoden unterstützt, welche mit Hilfe von ein- und zweisprachigen Korpora und bereits bestehenden Segmentpaaren der Übersetzungsspeichersysteme immer bessere (Fach)Übersetzungvorschläge und Items präsentierten. Dieses Herangehen beeinflusste in den letzten Jahrzehnten die Entstehung von zahlreichen neuen Korpora. Nunmehr ist die zweckorientierte Verarbeitung von Sprachdaten durch das Vorhandensein von Megakorpora der dritten Generation charakterisiert (Bank of English, Corpus of Contemporary American English).

Korpora spielen daher eine immer bedeutendere Rolle bei der computergestützten (Fach)Übersetzung (Machine-Aided Human Translation). Sie werden einerseits als online erreichbare Hilfsmittel beim Übersetzen verwendet und andererseits als Endprodukte der übersetzerischen Tätigkeiten erzeugt. Die Berufsübersetzer/innen und Übersetzerbüros verfügen bereits über verschiedene themenspezifische Paralleltexte, welche als Hintergrundkorpora der weiteren (Fach)Übersetzungen dienen können. Ihr Gebrauch ist zum Beispiel unerlässlich für Softwarelokalisierungen, wofür die Benutzeroberfläche eines Software- oder Webproduktes und die dazugehörige Produktdokumentation häufig parallel zur Entwicklung des Quellproduktes erzeugt werden muss.

Softwarelokalisierungen und auch Hintergrundkorpora sind stark kulturell gebunden. Bei einem lokalisierten Produkt „sieht die Benutzeroberfläche so aus,

als sei sie ursprünglich für den Zielmarkt geschrieben und entwickelt worden. Für die einwandfreie Lokalisierung eines Softwareprodukts oder einer Webseite müssen neben der Sprache u. a. folgende Aspekte berücksichtigt werden: Maßeinheiten, Zahlenformate, Adressformate, Datums- und Uhrzeitformate (lang und kurz), Papierformate, Schriftarten, Auswahl der Standardschrift, unterschiedliche Groß- und Kleinschreibung, Zeichensetzung, Sortierung, Wort- und Silbentrennung, lokale Vorschriften, Urheberrechtsprobleme, Datenschutz, Zahlungsmethoden, Währungsumrechnung, Steuern" (SDL TranslationZone). Wie diese Auflistung zeigt, müssen beim Übersetzen und bei der Softwarelokalisierung sämtliche kulturellen Eigenschaften der Ausgangs- und Zielsprachen in Betracht gezogen werden. Diese Eigenschaften sind in den monolingualen Korpora sowie in den Paralleltexten gespeichert. Mit der Speicherung und Wiederverwendung derartiger kulturspezifischer Daten wird die Erhaltung kulturgebundener Eigenschaften der Sprachen gesichert. Dieses Verfahren gewinnt aufgrund des potenziell wachsenden Einflusses des Englischen auf nahezu alle anderen Sprachen eine immer stärkere Bedeutung.

3. Forschungsinteresse der Übersetzungswissenschaft

Von den elektronisch gespeicherten Paralleltextpaaren können sogenannte Parallelkorpora (DeuCze-Korpus, QCRI AMARA Corpus, Slovak-Hungarian Parallel Corpus), die nicht nur für die computergestützte und maschinelle Übersetzung, sondern auch für die Übersetzungswissenschaft anwendbar sind, zusammengestellt werden. Im Interessenfeld der deskriptiven Übersetzungswissenschaft stehen sämtliche zielsprachliche Texteigenschaften, durch welche die von der muttersprachlichen Textproduktion auf lexikalischer, grammatischer oder textueller Ebene auftretenden Abweichungen aufgezeigt werden können (vgl. Seidl-Péch 2016). In den letzten Jahren befassten sich im Sinne dieser komparativen Betrachtungsweise zahlreiche Promotionsprojekte an der Eötvös-Loránd-Universität (Budapest/Ungarn) mit korpusgestützten Untersuchungen (Seidl-Péch 2011; Polcz 2012; Lengyel 2013; Mohácsi-Gorove 2014; Robin 2014; Sato 2014; Somodi 2014; Kovács 2015; Makkos 2015; Nagy 2015; Szijj 2015).

Durch das Entstehen von Übersetzungs- und Vergleichskorpora kann ein steigendes Interesse für korpusgestützte Betrachtungen der Translate verzeichnet werden. Zahlreiche Studien analysierten Übersetzungen als eine Art der Textproduktion und identifizierten die besonderen Eigenschaften übersetzter Texte. So entstanden (1) Übersetzungskorpora (translational corpus) aus Übersetzungen in eine Sprache und aus den dazugehörigen Originaltexten (English-Chinese Parallel Corpus) und (2) Vergleichskorpora (comparable corpus) aus Übersetzungen in eine Sprache und aus den original verfassten Texten derselben Sprache (z. B. Internati-

onal Corpus of English). Das ungarische Pannonia Corpus (Robin et al. 2016) enthält ein Übersetzungs-, wie auch ein Vergleichssubkorpus. Nach Zanettins Artikel über das English-Italian Translational Corpus (Zanettin 2002) bereichern diese Korpora unser Wissen über Sprachen, Kulturen und das Übersetzen. Heutzutage rücken weitere neue Themen in Fokus der übersetzungswissenschaftlichen Untersuchungen. Kürzlich erschienen u. a. korpusgestützte Studien zum Übersetzungsspeicher (Yamada 2011) oder zur Übersetzung von Dialekten (AsiCa-Korpus, Haddow et al 2013).

4. Kulturelle und wissenschaftliche Bedürfnisbefriedigung

In den letzten Jahrzehnten wird trotz Maßnahmen auf politischer Ebene im Dienste des Schutzes der kulturellen Vielfalt von einer Globalisierung und Entdifferenzierung im kulturellen Sinne gesprochen. Bisher wurde keine befriedigende Strategie zur kulturellen Nachhaltigkeit von allen Seiten akzeptiert, daher muss sich die globale Gemeinschaft zwangsläufig an irreversible kulturelle Verluste gewöhnen.

Multikulturelle und -ethnische Gesellschaften manifestieren sich in einer Vielfalt der Sprachen und Kulturen, die durch Sammlung von Sprachdaten und kulturell spezifischer Ausdrucksformen bewahrt werden kann. Korpora, als Datenspeichersysteme, können eine bedeutende Rolle hinsichtlich der kulturellen Nachhaltigkeit spielen. Sie tragen in einem Zeitalter, in dem kulturelle Entwicklungsprozesse weder blockiert noch verändert oder gesteuert werden können, zur Bewahrung der kulturellen Vielfalt bei.

Je nach Zielsetzung der empirischen Sprachbeschreibung oder des Anwendungszwecks enthalten Korpora verschiedene Datentypen. Die verschiedenen Korpustypen hinsichtlich der Anzahl der verarbeiteten Sprachen (einsprachige Korpora / bi- und multilinguale Korpora: Parallelkorpora, Übersetzungskorpora, Vergleichskorpora) wurden bereits (vgl. Kapitel 2. und 3.) beschrieben. Je nach Sprachmedium kann zwischen Korpora der gesprochenen Sprache (BEA Hungarian spontaneous speech database, Datenbank für Gesprochenes Deutsch, Hungarian Reference Speech Database, MTBA – Hungarian Phone Speech Call Center Database) und Korpora der geschriebenen Sprache (Deutsches Referenzkorpus – DeReKo, HG-1 Korpus, Szeged Treebank) unterschieden werden. In den letzten Jahren werden darüber hinaus multimodale Korpora (Sound Gesture Database – Hungarian) zusammengestellt, die extralinguistische Ereignisse, darunter sind videobasierte Aufzeichnungen wie Gesten, emotionale Ausdrücke oder Daten zu Blickbewegungen beim Lesen zu verstehen, enthalten. Unter multimodalen Kor-

pora sind Korpora der Gebärdensprachen (American Sign Language Linguistic Research Project Corpus, DGS-Korpus, First Hungarian Sign Language Corpus) besonders bemerkenswert. Die bisher uneinheitliche Annotation heterogener Daten bei multimodalen Korpora und bei Korpora der Gebärdensprachen stellt ein schwer zu überwindendes Hindernis dar. In dieser Hinsicht besteht innerhalb der Disziplin noch großer Entwicklungsbedarf.

Korpora werden zudem durch das Sprachstadium dichotomisiert: zu unterscheiden sind historische Korpora (Hungarian Generative Diachronic Syntax) und Korpora der Gegenwartssprache (Hungarian National Corpus). In diesem Zusammenhang entstehen und entstanden unter anderem Korpora schon ausgestorbener Sprachen (Electronic Text Corpus of Sumerian Literature) oder bedrohter Sprachen (Korpus GENIE – GEsprochenes NIEdersorbisch/Wendisch). Neben der Wahl der regionalen/nationalen, medialen und temporalen sprachlichen Varietäten und Varianten können natürlich noch weitere Korpusparameter, die die Zusammenstellung der Korpora bestimmen (wie Alter, Geschlecht und soziale Herkunft der Sprecher, Textsorte, mediale Verarbeitung, Themeninhalte, Zielgruppe...) in Betracht gezogen werden, diese sind jedoch in Bezug auf diese Überlegungen zu vernachlässigen und werden daher nicht berücksichtigt.

5. Schlussfolgerungen

Zwar wurden Korpora ursprünglich nicht im Rahmen einer Kulturschutzaktion erstellt, sie können jedoch durch ihre Datenspeicherfunktion der kulturellen Nachhaltigkeit und der Bewahrung der kulturellen Vielfalt dienen. Die derzeit erreichbaren Korpora erfüllen bereits eine Kulturschutzfunktion und mit dem Herstellen von zusätzlichen Korpora werden weitere Schritte zum Schutz der kulturellen Vielfalt realisiert.

Während in den Anfängen der wissenschaftlichen Dokumentation und Datensammlung das Erstellen von Korpora noch buchstäbliche Feldforschung beinhaltete, sind in den Zeiten der Digitalisierung die Zugänge zu Sprachdaten wesentlich erleichtert. Diskutiert werden kann darüber hinaus die Frage, ob das heute zugängliche Sprachmaterial authentischer und sprechernäher ist als Sprachaufzeichnungen unter Laborbedingungen. Gerade für Sprecher kleiner Sprachen oder Mitglieder kleinerer Kulturkreise bietet die Digitalisierung und Speicherung der Sprache und Kommunikation eine leicht realisierbare Methode der Form- und Funktionswahrung sprachlicher Charakteristika. Wissenschaftliche Auseinandersetzung und der Zugang zu jenen sprachlichen Äußerungen, kulturellen Phänomenen und Sprecherdaten sind nicht nur aufgrund von wissenschaftlicher Dokumentation, sondern auch mit Blick auf Erhalt und Schutz sprachlicher und kultureller Vielfalt eine Verpflichtung dieser Disziplin.

Literaturverzeichnis

Baker, M. 1993. Corpus Linguistics and Translation Studies. Implications and Applications. In: Baker., M.–Francis, G.–Tognini-Bonelli, E. (eds.): *Text and Technology: In Honour of John Sinclair.* Amsterdam–Philadelphia: John Benjamins Publishing Company, 233–250.

Francis, W. N.–Kučera, H. 1979. *Brown Corpus Manual.* Revised edition. Brown University. http://khnt.hit.uib.no/icame/manuals/brown

Glück, H. (Hrsg.) 2005. Metzler-Lexikon Sprache. Stuttgart–Weimar: Verlag J. B. Metzler.

Haddow, B.–Huerta, A. H.–Neubarth, F.–Trost, H. 2013. *Corpus development for machine translation between standard and dialectal varieties*. Proceedings of the Adaptation of Language Resources and Tools for Closely Related Languages and Language Variants. Bulgaria, 7–14.

Kovács, M. 2015. *Les aspects de traduction et de transmission de messages des phrasèmes universels dans le contexte de l'union européenne.* Erreichbar: http://doktori.btk.elte.hu/lingv/kovacsmarietta/thesis.pdf

Lemnitzer, L.–Zinsmeister, H. 2006. *Korpuslinguistik. Eine Einführung.* Tübingen: Narr.

Lengyel, I. 2013. *The Concept of Translation Error from a Functional Perspective.* Erreichbar: http://doktori.btk.elte.hu/lingv/lengyelistvan/thesis.pdf

Makkos, A. 2015. *Comparable Competencies in Mother Tongue and Translated Texts.* Links between competence in composition writing and competence in translating, based on a comparison of Hungarian texts produced by university students. Erreichbar: http://doktori.btk.elte.hu/lingv/makkosaniko/thesis.pdf

Mohácsi-Gorove, A. 2014. *The Concept of Translation Quality in Translation Studies.* Revision or Reviewing as a Guarantee of Quality Assurance. Erreichbar: http://doktori.btk.elte.hu/lingv/mohacsigoroveanna/thesis.pdf

Nagy, J. 2015. *The (Relative) Balance of Communicative Dynamism in Translation.* Erreicbar: https://edit.elte.hu/xmlui/bitstream/handle/10831/32440/thesis_eng_Nagy_Janos.pdf?sequence=4&isAllowed=y

Polcz, K. 2012. *Conventionally Indirect Speech Acts in English–Hungarian Film Script Translation.* Erreichbar: http://doktori.btk.elte.hu/lingv/polczkaroly/thesis.pdf

Robin, E. 2014. *Translation Universals in Revised Texts.* Erreichbar: http://doktori.btk.elte.hu/lingv/robinedina/thesis.pdf

Robin, E. et al. 2016. Fordítástudomány és korpuszkutatás: bemutatkozik a Pannonia Corpus. *Fordítástudomány* 18. évf. 2. szám. 5–26.

Sato, N. 2014. *The Dual Loyalty ft he In-House Dialogue Interpreter in Hungarian-Japanese and Japanese-Hungarian Interpersonal Communication.* Erreichbar: http://doktori.btk.elte.hu/lingv/satonoriko/thesis.pdf

SDL TranslationZone. Erreicbar: http://www.translationzone.com/de/solutions/software-localization/ (20.05.2017.)

Seidl-Péch, O. 2011. *Rechnergestüzter Vergleichanalyse von Zieltexten.* Vergleichanalyse vom lexikalischen Kohäsionsmuster der muttersprachlichen und der zielsprachlichen Textproduktion Ungarischer Texten. Erreichbar: http://doktori.btk.elte.hu/lingv/seidlpecholivia/thesis.pdf

Seidl-Péch O. 2016. Zu theoretischen und praktischen Aspekten des Fachübersetzens: Verwendbarkeit von Textkorpora für das Fachübersetzen und für die Übersetzungswissenschaft. *Acta Universitatis Sapientiae – Philologica* VIII. 3. 127–136.

Sinclair, J. 1991. *Corpus, concordance, collocation.* Oxford: Oxford University Press.

Somodi, J. 2014. *Pragmatics of Address Terms in Japanese-Hungarian Comparison Investigation of the Translation of Japanese Appellative Forms of Address in Hungarian Film Dialogue Texts.* Erreichbar: http://doktori.btk.elte.hu/lingv/somodijulia/thesis.pdf

Szijj, M. 2015. *Traducción literaria a lengua no materna.* Húngaros en la traducción de la literatura húngara al castellano. Erreichbar: http://doktori.btk.elte.hu/lingv/szijjmaria/thesis.pdf

Yamada, M. 2011. The effect of translation memory databases on productivity. In: Pym, A. (ed.): *Translation research projects 3*. Tarragona: Universitat Rovira i Virgili.

Zanettin, F. 2002. CEXI: Designing an English Italian Translational Corpus. In: Ketteman, B.–Marko, G. (eds.): 2002. *Teaching and Learning by Doing Corpus Analysis.* Amsterdam: Rodopi, 329–343.

Korpora

Albanian National Corpus	http://web-corpora.net/AlbanianCorpus/search/
Aligned Hansards ft he 36th Parliament of Canada	http://www.isi.edu/natural-language/download/hansard/
American Sign Language Linguistic Research Project Corpus	http://www.bu.edu/asllrp/
AsiCa-Korpus	http://www.asica.gwi.uni-muenchen.de
Bank of English	http://www2.lingsoft.fi/doc/engcg/Bank-of-English.html
BEA Hungarian spontaneous speech database	http://www.nytud.hu/adatb/bea/index.html
British National Corpus	http://www.natcorp.ox.ac.uk

Brown Corpus	http://www.helsinki.fi/varieng/CoRD/corpora/BROWN/
Bulgarian National Corpus	http://dcl.bas.bg/bulnc/en/
Czech National Corpus	https://www.korpus.cz
COMPARA	http://www.linguateca.pt/COMPARA/
Corpus of Contemporary American English	http://corpus.byu.edu/coca/
Corpus del Español	http://www.corpusdelespanol.org
Datenbank für Gesprochenes Deutsch	http://dgd.ids-mannheim.de/dgd/pragdb.dgd_extern.welcome
DeuCze-Korpus	http://www.deucze.germanistik.uni-wuerzburg.de
Deutsches Referenzkorpus – DeReKo	http://www1.ids-mannheim.de/kl/projekte/korpora.html
DGS-Korpus	https://www.sign-lang.uni-hamburg.de/dgs-korpus/index.php/korpus.html
Electronic Text Corpus of Sumerian Literature	http://etcsl.orinst.ox.ac.uk/
English-Chinese Parallel Corpus	http://ec-concord.ied.edu.hk/paraconc/index.htm
English-Norwegian Parallel Corpus	http://www.helsinki.fi/varieng/CoRD/corpora/ENPC/
European Parliament Proceedings 1996-2001	http://www.statmt.org/europarl/
First Hungarian Sign Language Corpus	http://jelesely.hu/web/?q=en
HG-1 Korpus	http://corpus.hungram.unideb.hu/
Hungarian Generative Diachronic Syntax	http://www.nytud.hu/depts/corpus/mgtsz.html
Hungarian National Corpus	http://mnsz.nytud.hu
Hungarian Reference Speech Database	http://alpha.tmit.bme.hu/speech/hdbMRBA.php
International Corpus of English	http://www.ucl.ac.uk/english-usage/projects/ice.htm
Korpus GENIE – GEsprochenes NIEdersorbisch/Wendisch	http://genie.coli.uni-saarland.de
Lancaster Corpus of Mandarin Chinese	http://www.lancaster.ac.uk/fass/projects/corpus/LCMC/
MTBA – Hungarian Phone Speech Call Center Database	http://alpha.tmit.bme.hu/speech/hdbMTBA.php
National Corpus of Polish	http://nkjp.pl/index.php?page=0&lang=1
QCRI AMARA Corpus	http://alt.qcri.org/resources/qedcorpus/

Quranic Arabic Corpus	http://corpus.quran.com
Romanian Balanced Corpus	http://metashare.elda.org/repository/browse/romanian-balanced-corpus-rombac/0a7dd85edc7311e5aa0b00237df3e35873a0d662435d42dd94fba48c29dc0065/
Russian National Corpus	http://www.ruscorpora.ru/en/index.html
Slovak-Hungarian Parallel Corpus	http://www.nytud.hu/depts/corpus/szlovak-magyar.html
Slovak National Corpus	http://korpus.juls.savba.sk/index_en.html
Sound Gesture Database – Hungarian	http://alpha.tmit.bme.hu/speech/gestures.php
Szeged Treebank	http://rgai.inf.u-szeged.hu/index.php?lang=en&page=SzegedTreebank
TIGER Corpus	http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger.html
Turkish National Corpus	http://www.tnc.org.tr