Grundlagen der Korpusarbeit

Was ist eigentlich ein Korpus?

Korpora sind Textsammlungen, die als Datengrundlage für die Untersuchung sprachlicher Phänomene dienen. Sie sind computerlesbar und häufig sehr umfangreich, was die maschinelle Auswertung und statistische Erfassung bestimmter Wörter, ihrer Flexionsformen oder ihrer Verwendungskontexte enorm erleichtert.

Jedes Korpus deckt einen ganz konkreten Referenzbereich ab und kann dadurch zur Beantwortung bestimmter sprachwissenschaftlicher Untersuchungsfragen genutzt werden. Das sehr bekannte FRANTEXT-Korpus beispielsweise umfasst 4500 vorwiegend literarische Texte des Französischen vom 12. Bis zum 21. Jahrhundert. Es eignet sich zur Untersuchung der diachronischen Entwicklung der geschriebenen Sprache oder zur Charakterisierung sprachlicher Besonderheiten in bestimmten literarischen Strömungen. Genauso gibt es aber auch Korpora zur gesprochenen Sprache in Form transkribierter Interviews (ESLO, CLAPI), Korpora zur Sprache im Internet (TWITA, WaCky), zur Jugendsprache (COLA) usw. Bevor man ein sprachliches Phänomen untersucht, sollte man sich stets fragen, ob das ausgewählte Korpus die nötige Repräsentativität für die Forschungsfrage aufweist. Um die Repräsentativität zu erhöhen, kann man bestehende Korpora auf Teilbereiche einschränken oder selbst Daten für ein Arbeitskorpus erheben.

Korpora sind stets an texttypologischen Kriterien orientiert, das heißt, sie umfassen Exemplare einer oder mehrerer genau definierter Textsorten, z.B. Literatur (oder konkreter: Mittelalterliche Ritterepen, spanischsprachige Romane des 20. Jahrhunderts), Zeitungen, Gesetztestexte, Blogs, Twitter-Mitteilungen, transkribierte Sprachaufnahmen (Dialoge, Interviews, Telefonate, Radiosendungen…). Sie werden entweder mit dem Ziel zusammengestellt, einen breiten Überblick über ein Genre zu vermitteln, oder bereits mit der Absicht, eine bestimmte Forschungsfrage zu beantworten. Im Internet (oder auch auf CDRom/DVD) werden sie schließlich einer großen forschenden Allgemeinheit zur weiteren Ergebnisgewinnung zur Verfügung gestellt.

Wie ist ein Korpus aufgebaut?

Um sie verlässlich durchsuchbar und im Internet verfügbar zu machen, werden Korpora in adäquaten Dateiformaten aufbereitet. Häufig wird hierbei XML (Extensible Markup Language) verwendet, eine Auszeichnungssprache, die mit einer Vielzahl von Textverarbeitungsprogrammen (z.B. Transcriber, Oxygen, TextPad, SCP, R) kompatibel ist und somit eine weitere Aufbereitung der Texte sowie die Suche nach bestimmten sprachlichen Phänomenen mit und ohne Suchmaske ermöglicht. XML ermöglicht auch die Integration von Metadaten in ein Textdokument. Hierzu zählen Informationen zu Sprache, Varietät und Textsorte, zum Autor bzw. Sprecher und dessen Alter, Geschlecht und sozialem Status, zum Zeitpunkt der Niederschrift/Veröffentlichung /Aufnahme der Daten sowie ggf. zu den Modalitäten der Transkription. All diese Informationen sind unabdingbar für die Kontextualisierung der Daten und die Klärung ihrer Repräsentativität für zu beantwortende Forschungsfragen.

Viele Korpora sind in Bezug auf Wortarten und Flexionsformen (Parts of Speech – PoS), Lemmata oder auch syntaktische Kategorien annotiert, was das Spektrum der möglichen Suchabfragen enorm erweitert. Dies unterscheidet sie von Textdatenbanken, die lediglich Texte in „Rohform“ zur Verfügung stellen. Textdatenbanken dienen eher dokumentarischen als sprachwissenschaftlichen Zwecken, durchaus können sie aber die Basis für ein selbst erstelltes und zu Analysezwecken weiter aufbereitetes Korpus darstellen.

Was kann man mithilfe von Korpora untersuchen?

Die Lemmatisierung und morphosyntaktische Annotation eines Korpus erlauben die Formulierung komplexer Suchanfragen unter Anwendung von Wildcards und regulären Ausdrücken, die Erstellung von Frequenzlisten (z.B. zur Ermittlung der Type-Token-Relation) und die Untersuchung von Konkordanzen zwischen Lemmata oder Wortarten. Auf diese Weise können charakteristische (lexikalische, syntaktische, pragmatische, phonetische…) Merkmale einer Sprache sowie Veränderungen derselben sichtbar und zählbar gemacht werden. Korpusanalysen spielen daher z.B. beim Sprachvergleich, bei der Beschreibung regionaler Varietäten, bei der Untersuchung von Sprachkontakt und Sprachwandel eine wichtige Rolle. Auch in der Soziolinguistik, welche sich mit der Analyse von Zusammenhängen zwischen sprachlichen und außersprachlichen Merkmalen (Alter, Bildung, soziale Stellung etc.) beschäftigt, finden sie Anwendung.

Die Arbeit mit Textdaten - Korpora und Textdatenbanken

Es lässt sich festhalten:
Ein linguistisches Korpus ist eine Textsammlung, die

- einen Referenzbereich systematisch abdeckt (z.B. Pariser Jugendsprache, diplomatische Korrespondenzen im Spanien des 18. Jh.).
- zur Beantwortung ganz bestimmter sprachwissenschaftlicher Untersuchungsfragen dient (z.B. Suche nach Anglizismen, Häufigkeit bestimmter Verbformen).

Ein Korpus muss quantitativ ausbalanciert sein. Das bedeutet nicht nur, dass die Textdaten einen gewissen Umfang aufweisen müssen, sondern auch, dass sie in ihrer Auswahl repräsentativ für den Referenzbereich sein müssen. Repräsentativität ist beispielsweise nicht gegeben, wenn man Aussagen über lexikalische Besonderheiten Südfrankreichs treffen will, jedoch ausschließlich Sprachdaten jüngerer Sprecher erhebt, da ihr Sprachgebrauch von dem älterer Sprecher abweichen kann und somit nicht mit dem generellen Usus Südfrankreichs gleichzusetzen ist.

Liegt der notwendige Datenumfang nicht vor, so sollte der Untersuchungsgegenstand, um eine hohe Repräsentativität zu erhalten, möglichst differenziert gewählt werden (z.B. lexikalische Besonderheiten der in Marseille lebenden Gymnasiasten). Je stärker die Fragestellung diatopisch, diastratisch und diaphasisch präzisiert wird, desto repräsentativer kann die Datengrundlage gewählt werden.

Auch sollte darauf geachtet werden, texttypologische Grenzen nicht zu überschreiten. In der Regel begrenzt sich die Datengrundlage auf einen einzigen Texttypus, z.B. Literatur, Zeitung, Gesetzestexte, Blogs, transkribierte Sprachaufnahmen, etc..

Im Hinblick auf das Datenformat muss vermerkt werden, dass Textdaten nicht einfach als Fließtext vorliegen, sondern in irgendeiner Weise aufbereitet wurden:

- Ihr Format orientiert sich an bestimmten Kodierungsstandards (z.B. XML), wodurch sie computerlesbar werden. Analyseprogramme und Suchmasken können anhand eines klar definierten Formats des Volltexts zielgerichtet Informationen hinzufügen, ändern, löschen oder herausfiltern.
- Sie sind annotiert, d. h. mit zusätzlichen (unsichtbaren) Informationen angereichert. Dabei kann es sich um Metadaten (z.B. Autor/ Sprecher, Alter, Ort, Datum, Varietät…) oder auch um linguistische Informationen (Wortarten, Lemmata) handeln.

Ein Kodierungsstandard macht es möglich, große Datenmengen automatisiert und schnell zu durchsuchen und zu analysieren, da er bewirkt, dass sprachliche Merkmale computerlesbar und dadurch zählbar gemacht werden. Sind die Textdaten mit linguistisch orientierten Annotationen oder Metadaten versehen, so können Sprachregister, Varietäten, Sprachwandel, Sprachkontakt usw. bei der Analyse der Frequenzen charakteristischer Merkmale berücksichtigt werden. Auf diese Weise wird dem Zusammenhang zwischen außersprachlichen Faktoren (Alter, Herkunft, Bildung…) und sprachlichen Phänomenen (Soziolinguistik) die notwendige Aufmerksamkeit zuteil.

Korpora im TdR-Wiki

Die Korpora lassen sich im TdR-Wiki über den Menüpunkt "Korpora und Textdatenbanken" annavigieren. Mit Ausnahme der offline verfügbaren Korpora, die eine eigenständige Kategorie bilden, sind sie nach Sprachen kategorisiert. Seiten zu den einzelnen Korpora enthalten zum einen eine Kurzdarstellung ihrer Zusammensetzung, zum anderen einen Link, der jeweils auf die entsprechende Wiki-externe Korpusseite führt. Korpora können zugansbeschränkt und somit nur über eine Universitäts-IP aufrufbar sein. Private Rechner können diese Beschränkung entweder durch das eduroam-Netz oder mithilfe eines VPN-Tunnels gegebenenfalls umgehen.

Einige Grundbegriffe der Korpusanalyse

Token - Type

Token: Wortformen (in einem definierten Text(abschnitt))
Types: unterschiedliche Wortformen (in einem definierten Text(abschnitt))
Bsp: (soy, soy, eres, hago)|(suis, suis, es, fais)|(sono, sono, sei, faccio) 4 Token, 3 Types

Lemma

Lemma: Die Grundform eines Wortes (in der Regel Singular für Nomina, Maskulinum SIngular für Adjektive)

Part of Speech (POS)

Wortart (morphosyntaktische Annotation)

Konkordanz, Kookkurrenz, KWIC (key word in kontext)

Konkordanz: Die sprachliche Umgebung eines Schlüsselwortes (automatisch hergestellt mit Konkordanzprogrammen)
Kookkurrenz: Wörter, die (gehäuft) gemeinsam auftreten
KWIC: Darstellungsformat von Konkordanzen eines Schlüsselwortes (z.B. *parce que*, siehe Bild)

cf. hierzu auch: Gerstenberg, Annette (2013): *Arbeitstechniken für Romanisten. Eine Anleitung für den Bereich Linguistik*. Berlin/Boston: De Gruyter, Kap. 6: "Benutzung von Korpora und Datenbanken", und: das Glossar von Sketch Engine.

Weitere hilfreiche Quellen zum Thema Korpora und Korpuslinguistik sind im Bereich Literatur zur Korpuslinguistik zu finden