Korpusanalyse in der computergestützten Komparatistik

,
christine_ivanovicl@hotmail.com
Universität Wien, Österreich

,
frank@geoinfo.tuwien.ac.at
Technische Universität Wien

Aufgaben der Vergleichenden Literaturwissenschaft (Komparatistik)

Komparatistische Forschung zielt weniger auf hermeneutische Auslegung von Einzeltexten, als darauf, (a) generalisierende Aussagen über (literarische) Texte, ihre Fomen und Funktionen zu machen, (b) deren historische Entwicklung innerhalb oder (c) im Austausch von kulturellen Systemen zu rekonstruieren, und (d) literarische Repräsentationen von 'Welterfahrung' mit anderen Repräsentationssystemen zu vergleichen.

Zu diesem Zweck vergleicht die Komparatistik eine Vielzahl von Texten (resp. von Texten und anderen künstlerischen Repräsentationsformen) in unterschiedlichen Sprachen. Vergleiche erfordern die Annahme einer Anzahl von Eigenschaften der verglichenen Gegenstände als gleichwertig, während andere Eigenschaften desselben Gegenstands variieren. Die Bestimmung der Eigenschaften eines Textes ist demnach eine der unabdinglichen Voraussetzungen für den komparatistischen Vergleich.

Korpusbasierte Komparatistik

Um allgemeine Aussagen machen zu können, muss die komparatistische Forschung andererseits eine größere Anzahl von Texten untersuchen: sie muss Textkorpora bilden und evaluieren. Die Anzahl und Auswahl der in Betracht gezogenen Texte sowie der evaluierten Texteigenschaften bestimmen maßgeblich die Ergebnisse einer komparatistischen Untersuchung.

Eine computergestützte korpusbasierte komparatistische Untersuchung unterscheidet sich von den bisher praktizierten Ansätzen nicht in der – die Disziplin charakterisierenden – Operation des Vergleichens, wohl aber in der Art und Weise, wie Auswahl und Anzahl der verglichenen Texte resp. Textkorpora begründet und dokumentiert werden.

Computergestützte korpusbasierte Komparatistik

Potentiell sind alle jemals verfassten und mündlich oder schriftlich tradierten Texte aller Sprachen und aller Zeiten Gegenstand der Komparatistik. Ein umfassender systematischer Zugriff auf alle diese Texte ist (bisher) jedoch nicht möglich. Die Möglichkeiten der Evaluierung sind durch verschiedene Faktoren limitiert: nicht alle Texte sind faktisch (mehr) zugänglich, und die jeweilige Forscherperspektive beschränkt grundsätzlich die Erfassung der zum Vergleich herangezogenen Texte.

Bisher sind die Kriterien für die Textauswahl in wesentlichem Maße abhängig gewesen von (a) der Subjektivität und (b) der natürlicherweise begrenzten Kapazitätder Forscher, die nur die ihnen bekannten Texte berücksichtigen können und die unter dem Credo arbeiten, nur Texte zu erforschen, die ihnen in der Originalfassung zugänglich sind. Dies führt dazu, dass die Komparatistik bisher mehrheitlich Texte aus den dominanten Sprachen (Englisch, Französisch,...) bearbeitet und Texte in 'kleinen' Sprachen (Finnisch, Urdu,....) oder Textvergleiche zwischen kaum verwandten Sprachen (Chinesisch gegen Arabisch) eher selten vorkommen.

Ein weiteres Problem der Textauswahl stellt (c) die Gefahr des logischen Zirkelschlusses dar: Bei der Evaluation beispielsweise “des” europäischen Romans werden aus der Lesepraxis resp. -tradition herrührende Vorannahmen in die Auswahl einbezogen, wenn es darum geht, dieses Genre anhand verschiedener Beispiele zu bestimmen; sie haben unweigerlich Einfluss auf das erzielte Ergebnis. Schließlich beruhen, und auch dies bedeutet eine wesentliche Einschränkung, (d) generalisierende Aussagen wie über “den europäischen Roman” immer auf einer im Vergleich zur Gesamtmenge der je produzierten Texte verschwindend kleinen Auswahl.

Die Auswahl der evaluierten Texte kann bei einer computergestützten korpusorientierten komparatistischen Untersuchung zumindest statistisch anders begründet werden:

* durch einen definitiv bestimmten Korpus, der so angelegt ist, dass er Repräsentativität beanspruchen kann
* durch einen Korpus, der in seinem Umfang weit über das Lesevermögen des Einzelnen hinausreicht und der große Textmengen in einer Vielfalt von Sprachen umfaßt, die kein Einzelleser je bewältigen könnte;
* durch die Möglichkeit der Überprüfung der erzielten Ergebnisse in Wiederholungs- und Vergleichsstudien sowie mittles Vergleichskorpora;
* durch die Trennung der Auswahlkriterien für die Erstellung des Korpus von den fokussierten Untersuchungsergebnissen;

* durch die Möglichkeit von Negativabfragen (z. B. eine bestimmte Eigenschaft ist in einigen der Texte des Korpus nachweisbar, während andere Eigenschaften in keinem davon nachweisbar sind).

Anforderungen an die computergestützte korpusbasierte Komparatistik

Computergestützt korpusbasiert arbeitende Komparatistik sieht sich mit folgenden Aufgaben konfrontiert:

Erarbeitung einer effizienten Infrastruktur

Für den Aufbau und die Pflege großer Textkorpora bedarf es entsprechend bearbeiteter Texte: alle in den Korpus aufgenommene Texte müssen bibliographisch genau erfasst und mit Markups (Taggern) versehen sein. Markups können gesetzt werden u. a. zur Auszeichnung der Sprachform (insbesondere bei mehrspachigen Texten), der Textstruktur, nicht-literarischer Elemente (z. B. Abbildungen im Text) etc. Bevorzugt werden treebank getaggte Versionen mit verzweiger Struktur (parse tree), die Koreferenzen, Personen- und Orstnamen u. a. erkennen lassen. Im Textmarkup werden einzelne Elemente der Textstruktur identifiziert: Worte oder Wortbestandteile, Sätze und deren Teile, Abschnitte, Kapitel und andere Texteinteilungen bis hin zum Buchlayout. Es erscheint wichtig, auch die Elemente zu erfassen, die nicht unmittelbar textimmantent sind, die aber zur Identifizierung und Charakterisierung des Textes gehören wie Seitenangaben, Verfassername und weitere Angaben, die im Rahmen einer Buchpublikation vorkommen. Der Text sollte in UTF-8 codiert sein, um auch Texte in nicht-alphabetischen Sprachen wie Chinesisch, Arabisch u.w.m. einbeziehen zu können. Unserer Konzeption nach sollen die Markierungen in den Text hineingesetzt werden, so dass der mit den Annotationen versehene Text mit der Originalstruktur verbunden bleibt.

Der mit Markups versehene Text und die POS-Annotationen werden in ein einziges Format zusammengefasst. Wir bevorzugen derzeit RDF (Manola / Miller 2004), das für die von uns avisierte Datenmenge auszureichen scheint. Unseren bisherigen Beobachtungen nach erhalten wir bei einem Text mit reichhaltiger linguistischer Auszeichnung für jedes Wort etwa 10 RDF Triples; bei einem literarischen Text von 100.000 Wörtern würde das eine Million Triples ergeben, bei einem Korpus von 10.000 Büchern wäre man mit 10 Milliarden Triples noch bei weitem innerhalb des Rahmens dessen, was das heutige RDF Depot erlaubt; Untersuchungen zur derzeitigen Kapazitätsgrenze haben für 1 Billiarde Triples eine Hochladezeit von wenigen Stunden ergeben (Boncz / Pham 2013). Die Antwortzeiten nehmen bei unterschiedlichen Abfragen nicht ab und die Anforderungen der Hardware bleiben im Rahmen des für ein Projekt Möglichen (die Anschaffungskosten der Versuchskonfiguration von 2012 beliefen sich auf 70.000 Euro).

Entwicklung brauchbarer Methoden zur Abfrage und digitalen Analyse von Texten

Es müssen Methoden sein, die Abfragen und Analysen von Texten ermöglichen unabhängig von der Sprache, in der sie verfasst sind, d. h. wir arbeiten an (statistischen) Methoden, die Texteigenschaften in (mathematische) Werte 'übersetzen', um eine Vergleichbarkeit von Textstrukturen über die Sprachgrenzen hinweg zu ermöglichen. Die einschlägige Fachliteratur kennt bereits eine große Anzahl derartiger Methoden; auf viele von ihnen kann man über das web zugreifen. Die computergestützt korpusbasiert arbeitende Literaturwissenschaft erlaubt z. B. die Identifizierung von Zitaten und intertextuellen Bezugnahmen (Ganascia et al. 2014). Sie dient dem Autornachweis oder der Evaluierung von gender-bedingten Spezifika (Stanczyk 2011), wie auch der Feststellung literarischer Moden und Bewegungen (Amancioa et al. 2012). Sie unterstützt die Analyse von Emotionen (Dichiu et al. 2010), oder die Rekonstruktion von Wissenstransfer (Cappelli et al. 2001).

Einführung von Clusteranalyse in die Komparatistik

Die Anwendung aller Methoden auf alle Texte generiert eine Matrix evaluierbarer Werte; jeder Text läßt sich durch einen Vektor aus diesen Werten darstellen. Diese Darstellungsweise ermöglicht Textvergleiche mittels Clusteranalyse wie sie in der konventionellen Komparatistik aufgrund der o.g. Beschränkungen bisher nicht zugänglich waren.

Korpusaufbau und Abfragemethoden müssen so gestaltet sein, dass Texte umstandslos dem Korpus hinzugefügt werden und die Methoden problemlos appliziert werden können. Dies setzt kontinuierliche Pflege und Aktualisierung des bestehenden Korpus resp. der Abfrageergebnisse voraus: wenn Texte hinzugefügt werden, müssen alle bisher angewandten Methoden automatisch darauf angewandt werden können; wenn Methoden hinzugefügt werden, müssen automatisch alle Texte einer entsprechenden Evaluierung unterzogen werden.

Zusammenfassung

In unserem Beitrag treten wir für die Etablierung eines computergestützten korpusbasierten Forschungsansatzes

in der literaturwissenschaftlichen Komparatistik ein. Dazu wollen wir darstellen, (a) welche Vorteile die Erstellung umfangreicher Korpora literarischer Texte aus verschiedenen Sprachen für die komparatistische Analyse bietet, (b) wie sie konstruiert und gepflegt werden können, und (c) welche Abfragemöglichkeiten auf dem gegenwärtigen Stand der Technik sie bieten. In Betracht gezogen werden dafür sowohl bereits vorhandene und online zugängliche Korpora wie auch von einzelnen Forschergruppen erarbeitete, intern genutzte Korpora wie das Austrian Academy Corpus am ICLTT der ÖAW. Des weiteren wollen wir einen kursorischen Überblick über die bisher erprobten Ansätze computergestützter literaturwissenschaftlicher Analyse geben, um das gegenwärtige Spektrum der Methoden der Textevaluierung darstellen und zukünftige Desiderata aufzeigen zu können.

Bibliographie

Amancio, Diego R. /Oliveira Jr., Osvaldo N. / F. Costa, Luciano da (2012): "Identification of literary movements using complex networks to represent texts", in: *New Journal of Physics* 14 http://arxiv.org/abs/1302.4099 [letzter Zugriff 09. Januar 2016].

Boncz, Peter / Pham, Minh-Duc (2013): *BSBM V3.1 Results (April 2013)*. Centrum Wiskunde & Informatica, Amsterdam http://wifo5-03.informatik.uni-mannheim.de/bizer/berlinsparqlbenchmark/results/V7/ [letzter Zugriff 09. Januar 2016].

Cappelli, Amedeo / Catarsi, Maria Novella / Michelassi, Patrizia / Moretti, Lorenzo / Baglioni, Miriam / Turini, Franco / Tavoni, Mirko (2002): "Knowledge mining and discovery for searching in literary texts", in: *Proceedings of the Third International Conference on Language Resources and Evaluation*, LREC 2001, May 29-31, 2002, Las Palmas, Canary Islands, Spain.

Dichiu, Daniel /Pais, Ana Lucia / Moga, Sunita Andrea / Buiu, Catalin (2010): "A cognitive system for detecting emotions in literary texts and transposing them into drawings", in: *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*. Istanbul, Turkey, 10-13 October 2010: 1958-1965.

Ganascia, Jean-Gabriel / Glaudes, Pierre / Del Lungo, Andrea (2014): "Automatic detection of reuses and citations in literary texts", in: *Literary and Linguistic Computing* 29, 3: 412-421.

Manola, Frank / Miller, Eric (eds.) (2004): *RDF primer 1.0*. W3C recommendation, 10 (1-107): 6 https://www.w3.org/TR/2004/REC-rdf-primer-20040210/ [letzter Zugriff 09. Januar 2016].

Moretti, Franco (1999): *Atlas of the European novel*. 1800-1900. London, New York: Verso.

Stanczyk, Ursula (2011): "Recognition of author gender for literary texts", in: *Man-Machine Interactions* 2. Proceedings of the 2nd International Conference on Man-Machine Interactions, ICMMI 2011, The Beskids, Poland.

Takeda, Masayuki / Fukuda, Tomoko / Nanri, Ichiro (2002): "Mining from literary texts: Pattern discovery and similarity computation", in: *Progress in Discovery Science*. Final Reports of the Japanese Discovery Science Project. Berlin / Heidelberg: Springer 518-531.