Einführung

Medienlinguistische Methodik

Arne Rubehn
arne.rubehn@uni-passau.de

14.10.2025

1 Organisatorisches

14.10.	Einführung
21.10.	Modellierung
28.10.	Datenerhebung: Beobachtungen und Befragungen
04.11.	Datenerhebung: Experimente
11.11.	Korpuslinguistik
18.11.	Annotationen
25.11.	Textanalyse
02.12.	Diskursanalyse
09.12.	Daten aus Social Media
16.12.	Sprachmodelle
13.01.	Qualitative Inhaltsanalyse
20.01.	Offene Forschung
27.01.	Zusammenfassung & Klausurvorbereitung
03.02.	**Klausur (120 min. – MGP mit SE Werbesprache)**

Es besteht **keine Anwesenheitspflicht**. Es wird **keine zusätzliche Lektüre** erwartet. In den Handouts finden sich weitere Referenzen zu jedem Thema, die auf Anfrage gerne bereitgestellt werden können. Im Laufe des Semesters werden mindestens **drei kleine Studienleistungen** erwartet, die während den Sitzungen bearbeitet werden können.

2 Erste Denkanstöße

2.1 Die wissenschaftliche Methode

Wir starten dieses Seminar mit einem kleinen Exkurs in die Wissenschaftsphilosophie, also der Disziplin, die sich damit beschäftigt, was Wissenschaft überhaupt ausmacht und wie Wissen generiert wird. Als Teil dieser Disziplin hat sich die **wissenschaftliche Methode** im Zuge der europäischen Aufklärung herauskristallisiert, um wissenschaftliche Vorhaben von nicht-wissenschaftlichen abgrenzen zu können.

Abb. 1 zeigt eine schematische Darstellung der wissenschaftlichen Methode. Zu Beginn steht stets eine Beobachtung: Es geht also zunächst einmal darum, interessante Phänomene zu finden, die man untersuchen kann. Hat man ein solches Phänomen gefunden, beginnt man mit der Fragestellung, warum gewisse Muster denn vorkommen. Basierend auf diesen Fragen stellt man Hypothesen auf, womit gewisse Phänomene denn möglicherweise zu begründen seien. Um diese Hypothesen zu testen, müssen überprüfbare Vorhersagen abgeleitet werden – wenn die Hypothese korrekt ist, müsste X unter Bedingung Y eintreffen. Diese Vorhersagen werden dann schließlich überprüft, indem relevante Daten gesammelt und analysiert werden. Stellt sich die Hypothese als unzureichend her-

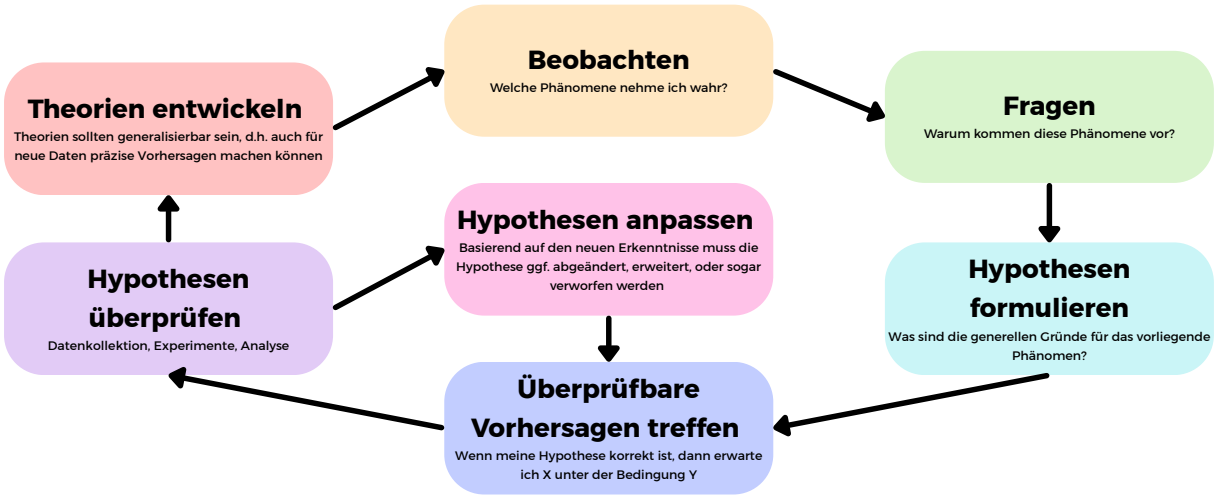


Abbildung 1: Illustration der wissenschaftlichen Methode (adaptiert von Green, o.D.).

aus, muss sie revidiert werden, also entweder angepasst, erweitert, oder gar komplett verworfen werden. In diesem Fall müssen wieder, basierend auf der angepassten Hypothese, neue Vorhersagen getroffen werden, die es wiederum zu überprüfen gilt. Ist die Hypothese andererseits korrekt, kann sie als Grundlage für generellere Theorien dienen, die den Anspruch haben, grundlegende Erklärungen für vergleichbare Muster zu finden. Dabei ist es wichtig, dass solche Theorien generalisierbar sein müssen; sie müssen also auch auf neue, aber vergleichbare Daten zutreffen und präzise Vorhersagen machen können.

Etwas verwunderlich scheint auf den ersten Blick der Pfeil, der von der Entwicklung der Theorien wieder zurück auf den ersten Schritt der Beobachtung zeigt. Jedoch ist es tatsächlich so, dass gute Theorien wertvolle Hinweise darauf geben können, **wohin** Forschende als nächstes schauen sollten. Ein sehr direktes Beispiel hierfür ist die Entdeckung des Planets Neptun: Diese wurde nämlich erst durch die Newtonsche Gravitationstheorie möglich! Isaac Newton wiederum entwickelte seine Theorie aus der Beobachtung heraus, dass sich Planeten nicht in perfekten Kreisen oder Ellipsen um die Sonne bewegen, sondern ihre Umlaufbahnen immer wieder leicht abweichen können. Das begründete er damit, dass auch Planeten untereinander sich gegenseitig anziehen können, wenn sich ihre Umlaufbahnen nahekommen – eine Vermutung, die die vermeintlichen Unregelmäßigkeiten der einzelnen Planeten schlüssig erklärte (und später auch experimentell nachgewiesen wurde). Lediglich beim Planeten Uranus – damals der äußerste bekannte Planet des Sonnensystems – beobachtete man nach wie vor Unregelmäßigkeiten, die die Gravitationstheorie basierend auf den aktuell bekannten Himmelskörpern nicht erklären konnte. Allerdings könnte man diese Bewegungen im Rahmen der Gravitationstheorie durchaus erklären, wenn es noch einen weiteren, bislang unentdeckten Planeten gäbe, der ebenso Anziehungskräfte auf Uranus auswirkte – dieser Vermutung folgend errechnete der französische Mathematiker Urbain Le Verrier die Position dieses hypothetischen Planeten. Seine Berechnungen schickte er an das Berliner Observatorium, wo Johann Gottfried Galle dann tatsächlich – fast genau an der von Le Verrier errechneten Position – einen neuen Him-

melskörper sichten konnte, der heute als Neptun bekannt ist (Kollerstrom, 2001).

So bildet sich also ein Zirkel, der untermalt, dass wissenschaftliche Praxis ein hochgradig iterativer Prozess ist. Im Rahmen der wissenschaftlichen Methode kann man zwischen vier grundlegenden Elementen unterscheiden, die für die Forschung notwendig sind: **Charakterisierungen** (Beobachtungen, Definitionen, Messungen), **Hypothesen** (Theoretische Erklärungen für die beobachteten Phänomene), **Vorhersagen** (Testbare Prozesse, die sich aus der Hypothese ableiten) und **Experimente** (Test der vorherigen drei Elemente). Es bleibt allerdings festzuhalten, dass diese Definition der wissenschaftlichen Methode wohl besser auf Forschungsrichtungen zutrifft, die mit kontrollierten Experimenten arbeiten kann. Insbesondere in den Geistes- und Sozialwissenschaften ist das häufig nicht möglich, wodurch Erkenntnisse eher „indirekt" aus Daten gewonnen werden müssen, was selbstverständlich auch das Prozedere etwas abändert.

Dennoch fußen auch diese Forschungsrichtungen auf den selben Grundsätzen, die elementar für wissenschaftliches Arbeiten sind: **Ehrlichkeit, Offenheit** und **Falsifizierbarkeit**, wobei der letzte Aspekt so zu verstehen ist, dass jede wissenschaftliche Theorie – basierend auf neuer Evidenz oder Methodik – prinzipiell widerlegt werden kann. Doch inwiefern ist Wissenschaft tatsächlich offen und ehrlich?

Zu Beginn des 21. Jahrhunderts erschienen einige Studien, die die Wiederholbarkeit von anderen Studien überprüfte. Dabei zeigten sich, insbesondere in den Fachbereichen der Medizin und der Psychologie, schockierende Ergebnisse: Ein Großteil der publizierten wissenschaftlichen Ergebnisse konnte nicht reproduziert werden! Ioannidis (2005) zeigt, dass es hierbei einen starken Zusammenhang zwischen Kompetitivität und Reproduzierbarkeit gibt: Je höher der wirtschaftliche oder akademische Druck ist, Ergebnisse zu publizieren, umso höher ist die Wahrscheinlichkeit, dass publizierte Ergebnisse nicht reproduzierbar sind. Diese Erkenntnisse gingen als **Replikationskrise** (*replication crisis*) in das Erbe der Wissenschaften ein und erzeugten viel Aufmerksamkeit. Immerhin scheint doch ein gewisses Umdenken stattgefunden zu haben: Neuere Studien zeigen, dass wissenschaftliche Untersuchungen seitdem tatsächlich besser reproduzierbar geworden sind (Korbmacher et al., 2023). Dies ist nicht zuletzt auch Verdienst der *Open Science*-Bewegung, deren Praxis wir auch im Rahmen dieses Seminars kennenlernen werden.

2.2 Semesterausblick

Das Seminar wird unter dem Hauptthema **Korpuslinguistik** stehen. Es wird also primär darum gehen, wie wir den Sprachgebrauch mithilfe von großen Textsammlungen untersuchen können. Des Weiteren sollen wissenschaftliche Prozesse im Allgemeinen diskutiert werden (wie im vorherigen Abschnitt schon angeschnitten).

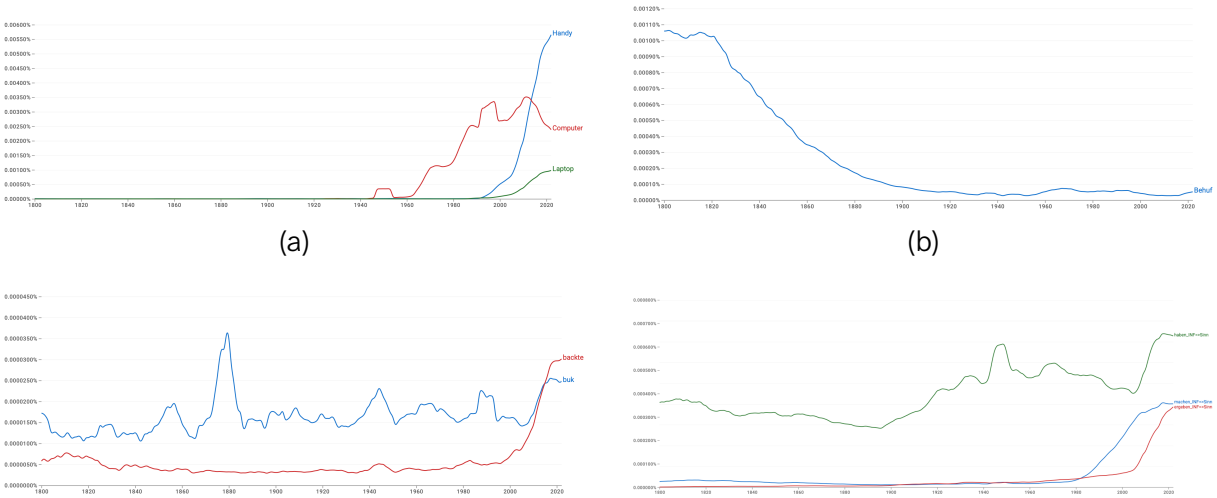
Hierbei wird zunächst einmal die wissenschaftliche Methodik per se thematisiert. Wir ler-

nen kennen, was wir uns unter Daten und Modellen in der Wissenschaft vorstellen können (Modellierung), wie wir Daten mithilfe von Beobachtungen, Befragungen oder Experimenten gewinnen können (Datenerhebung) und wie wir bestehende Daten zu Korpora zusammenfassen und als solche analysieren können (Korpuslinguistik). Des Weiteren wird illustriert, wie und zu welchem Zweck Daten annotiert werden können (Annotationen).

Der zweite größere Themenblock des Semesters widmet sich Techniken, mit denen Korpusdaten systematisch untersucht oder zu anderen Zwecken verwendet werden können. Hierbei stehen zunächst die Textanalyse und die Diskursanalyse im Mittelpunkt. Anschließend werden als „Sonderfall“ Daten aus Social Media besprochen, dann sehen wir uns an, wie solche Daten zum Training von großen Sprachmodellen wie (Chat-)GPT verwendet werden. Den Abschluss dieses Themenblockes macht dann die qualitative Inhaltsanalyse als komplementärer Ansatz zu den anderen, vorwiegend quantitativ ausgerichteten Methoden.

Zum Abschluss soll wieder der Bogen zur generellen Methodik in der Wissenschaft geschlagen werden, indem wir uns dem Thema der offenen Forschung widmen. Hierbei soll es sowohl um theoretische Grundsätze gehen, als auch um praktische Richtlinien, mit denen diese Grundsätze umsetzbar sind.

2.3 Google Books Ngram Viewer



(a) (b)

(c) (d)

Abbildung 2: Häufigkeiten von einigen Wörtern und Wortkombinationen zwischen 1800 und 2022.

Diese Sitzung beschließen soll ein kleiner Vorgeschmack auf das, wie große Textkorpora denn aussehen können und was wir mit ihnen machen können. Abb. 2 zeigt Screenshots aus dem *Google Books Ngram Viewer* (Michel et al., 2011), ein diachroner Korpus aus Büchern in verschiedenen Sprachen zwischen 1800 und 2022. Man sieht schön, welche

Information über den Sprachwandel in diesem Korpus steckt: Man kann nachvollziehen, welche Wörter wann neu in die Sprache Einzug erhalten haben (a), welche Wörter wann aus dem Wortschatz verschwunden sind (b), wie gewisse Flektionsformen über die Zeit grammatikalisiert werden (c) oder wie sich Kollokationsprofile über die Zeit wandeln (d).

Damit das möglich ist, müssen die Korpusdaten natürlich gründlich aufbereitet und mit weiteren Informationen angereichert werden. Zum Einen benötigt es maschinell lesbare **Primärdaten**, also die Texte selbst. Um diese zu erzeugen, sind Scans der Originalbücher nötig, die dann transkribiert werden müssen (üblicherweise mithilfe von Optical Character Recognition Tools, kurz OCR). Des Weiteren werden **Metadaten** benötigt, die jedes Buch beschreiben. Unabdingbar für die Funktionsweise des Ngram Viewers sind natürlich das Jahr und die Sprache, allerdings sind auch andere Informationen (z.B. Genre oder Autor:in) denkbar. Nicht zuletzt ermöglicht der Ngram Viewer auch komplexere Suchanfragen, sodass z.B. nach allen flektierten Formen eines bestimmten Wortes gesucht werden kann, oder nach der Kombination eines bestimmten Wortes mit einer Wortklasse (bspw. „Sinn + VERB“). Dies ist nur durch **Annotationen** möglich, die jedes Wort mit zusätzlichen Informationen wie Wortklasse, Lemma oder Bezugswort im Satz anreichern (Lin et al., 2012).

Viele Studien verwenden den Ngram Viewer, um die Entwicklung des Lexikons diachron zu untersuchen. Eine prominente Beispielstudie sei hier erwähnt, die den Bedeutungswandel von einzelnen Wörtern anhand von jeweiligen Kontextwörtern nachvollzieht (Hamilton, Leskovec & Jurafsky, 2016). Man sieht hier also schon deutlich: Große Textkorpora bergen ein unheimliches Potenzial, viele Aspekte der Sprache systematisch zu untersuchen.

Literatur

Green, W. (o.D.). *Scientific method*. Zugriff auf https://www.mrgscience.com/scientific-method.html (Aufgerufen am 13.10.2025)

Hamilton, W. L., Leskovec, J. & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th annual meeting of the acl (volume 1: Long papers)* (S. 1489-1501). Berlin, Germany: ACL. Zugriff auf https://www.aclweb.org/anthology/P16-1141 doi: 10.18653/v1/P16-1141

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine*, *19* (8), e1004085. doi: https://doi.org/10.1371/journal.pmed.1004085

Kollerstrom, N. (2001). *A neptune discovery chronology*. University College London. Zugriff auf https://web.archive.org/web/20051119031753/http://www.ucl.ac.uk/sts/nk/neptune/chron.htm (Aufgerufen am 13.10.2025)

Korbmacher, M., Azevedo, F., Pennington, C. R., Hartmann, H., Pownall, M., Schmidt, K., ... Evans, T. (2023). The replication crisis has led to positive structural, procedural, and

community changes. *Communications Psychology*, *1* (1). Zugriff auf http://dx.doi.org/10.1038/s44271-023-00003-2 doi: https://doi.org/10.1038/s44271-023-00003-2

Lin, Y., Michel, J.-B., Aiden Lieberman, E., Orwant, J., Brockman, W. & Petrov, S. (2012). Syntactic annotations for the Google Books NGram corpus. In *Proceedings of the ACL 2012 system demonstrations* (S. 169–174). Jeju Island, Korea: ACL. Zugriff auf https://aclanthology.org/P12-3029/

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Team, G. B., ... others (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331* (6014), 176–182.