Alexander Koplenig

Prinzipielle Probleme der Anwendung statistischer Signifikanztests in der Korpuslinguistik

> Wohl in kaum einem Bereich der Datenanalyse finden sich mehr Mißverständnisse, Fehlinterpretationen und Halbwahrheiten als bei der Anwendung und Interpretation von Signifikanztests, und zwar nicht nur bei Laien, sondern häufig auch bei gestandenen Fachleuten.
> (Diekmann, 2002, S. 585f)

In dem Vortrag sollen anhand von quantitativen Beispielen und computergestützten Simulationen einige Argumente vorgestellt werden, die dafür sprechen, dass die Anwendungsvoraussetzungen, welche dem Prinzip des statistischen Signifikanztests zugrunde liegen, d.i. der Schluss von den Eigenschaften einer Stichprobe auf die Eigenschaften einer Grundgesamtheit, aus prinzipiellen Gründen in der Korpuslinguistik – als wichtige Teildisziplin der Digitial Humanities – nicht erfüllt sind. Folgt man diesen Argumenten so ergeben sich für die Korpuslinguistik unter Umständen weitreichende Folgen.

In der Korpuslinguistik wird angenommen, dass die (relative Token-) Vorkommenshäufigkeit bestimmter sprachlicher Strukturen mit der kognitiven Repräsentation bzw. Prototypikalität dieser Strukturen verbunden ist, oder anders ausgedrückt, dass Korpushäufigkeit kognitive Verankerung instanziiert. Findet man dann zum Beispiel heraus, dass gewisse sprachliche Strukturen in einem Korpus geschriebener Sprache häufiger auftreten als in einem Korpus gesprochener Sprache, so folgert man daraus, dass dieser Struktur ein wichtigerer Status im geschriebenen Diskurs verglichen mit dem gesprochenen Diskurs zukommt (Schmid, 2010). Das Ziel einer korpuslinguistischen Studie ist es also weniger Aussagen über die untersuchten Korpora zu tätigen als vielmehr von diesen Textsammlungen auf die sprachlichen Varietäten zu schließen, die sie als Ausschnitt repräsentieren sollen (Baroni & Evert, 2009, S. 2).

In diesem Zusammenhang wird typischerweise ein statistischer Signifikanztest verwendet, um zu belegen, wie sicher man sich sein kann, dass der gefundene Zusammenhang nicht nur zufällig aufgetreten ist oder auch wie sicher man sich sein kann, dass die Hypothese „in Wahrheit“ wirklich richtig ist. Diese Annahme trifft aus statistischer Sicht jedoch nur unter ganz bestimmten Voraussetzungen zu.

Ein kurzer Exkurs in die Bevölkerungswissenschaften soll dies näher beleuchten: Bei Wahlumfragen geht es zum Beispiel darum, mit Hilfe der Befragung einer Auswahl von wahlberechtigten Personen auf das Ergebnis der Wahl schließen zu können. Die Stichprobe besteht dabei aus den befragten Personen, während sich die Grundgesamtheit aus allen wahlberechtigten Personen zusammensetzt. Mit Hilfe der Daten lässt sich dann beispielsweise ein Zusammenhang zwischen Wahlabsicht und Beruf berechnen.
Entscheidend ist hierbei, dass die Personen, die befragt werden, per Zufall ausgewählt werden. Nur dann lassen es die Grenzwertsätze der Statistik zu, Eigenschaften der Grundgesamtheit über die Stichprobe zu quantifizieren (Jann, 2005, S. 124–127). Statistische Signifikanz hat dabei nichts mit der Wichtigkeit des Forschungsergebnisses zu tun. Vielmehr liefert ein statistischer Test „eine formale Entscheidungsregel, die aufgrund einer Stichprobe darüber entscheidet, ob [der gefundene Zusammenhang] für die Grundgesamtheit zutrifft." (Fahrmeir, Künstler, Pigeot, & Tutz, 2001, S. 404). Ein Signifikanzniveau von 0,01 beruht auf folgendem Gedankenexperiment: angenommen man würde die Zufallsauswahl und Befragung der wahlberechtigten Personen sehr häufig wiederholen, dann würde sich das in der tatsächlich vorhandenen Befragung erzielte Ergebnis nur in höchstens 1 Prozent aller (hypothetischen) Stichprobenziehungen einstellen, obwohl es in Wirklichkeit, d.h. in der Grundgesamtheit, überhaupt nicht vorhanden ist.

Folgt man dieser Definition, so stellen sich für die Anwendung des Verfahrens in der Korpuslinguistik einige grundlegende Probleme: Ob sich ein Unterschied als statistisch signifikant erweist, hängt neben der absoluten Größe des Unterschieds vor allem von der Größe der Stichprobe ab. Das bedeutet, dass aufgrund ständig wachsender Korpusgrößen auch völlig unbedeutende Zusammenhänge signifikant werden.
Weiterhin gilt, dass man kein Korpus als eine (endliche) Zufallsstichprobe der ohnehin schwer definierbaren jeweiligen „Sprache als Ganzes" im strikten statistischen Sinn bezeichnen kann (Baroni & Evert, 2009, S. 3). Findet man im eingangs erwähnten Beispiel einen signifikanten Unterschied zwischen dem Korpus geschriebener Sprache und dem Korpus gesprochener Sprache, so könnte bei Wahl einer anderen Korpusgrundlage auch das Ergebnis völlig anders ausfallen, was besonders bei seltenen sprachlichen Phänomenen problematisch sein kann.
Dabei bietet die statistische Methodologie keinerlei Hilfestellung bei der Beantwortung der Frage, welche von beiden Untersuchungen denn nun eher der „Wahrheit" entspricht. Wenn jedoch verschiedene Untersuchungen mit unterschiedlichen Korpusgrundlagen in die gleiche Richtung

deuten, so kann man dies durchaus vorsichtig als Indikator für einen tatsächlich vorhandenen Unterschied zwischen den beiden sprachlichen Varietäten deuten.

Auf der anderen Seite wird dieses Problem zusätzlich dadurch verschärft, dass gewisse Arten von Texten (prinzipiell) nicht in einem Korpus erscheinen, man denke zum Beispiel an intime Gespräche zwischen Ehepartnern oder Diplomaten/-innen. Angenommen man würde solche Gespräche mit Einverständnis der Beteiligten aufnehmen, um sie anschließend in ein Korpus zu überführen, so ist davon auszugehen, dass die Aufnahme des Gesprächs reaktiv ist (Diekmann, 2002, S. 520–523), d.h. dass das eigentliche Gespräch durch die Messung beeinflusst wird. In Korpora geschriebener Sprache gilt dies in ähnlicher Weise für Werke, welche man aus Urheberrechtsgründen nicht veröffentlichen kann. Darüber hinaus werden ja gerade Zeitungstexte, die oftmals den Hauptbestandteil einer Textsammlung ausmachen, vor der Veröffentlichung nach bestimmten Regeln redigiert und können deshalb nur bedingt als prototypischer Ausschnitt der geschriebenen Sprache bezeichnet werden (Gries & Berez, noch nicht erschienen, S. 2).

Daher gilt, dass man ein Korpus als opportunistische Stichprobe der jeweils untersuchten Sprache bezeichnen muss (Lüdeling & Evert, 2005, S. 6). Gleichzeitig sei darauf hingewiesen, dass dies nur dann aus statistischer Sicht ein Problem ist, wenn die nicht vorhandenen Sprachbelege systematisch verzerrt sind, d.h., dass das untersuchte sprachliche Phänomen in den nicht vorhanden Texten anders repräsentiert ist (Diekmann, 2002, S. 357–359). Jedoch gibt es auch hier keine Methode, mit der man prüfen könnte, ob es sich um eine systematisch verzerrte Stichprobe handelt.

Nun könnte man einwenden, dass es in weiten Teilen der empirisch arbeitenden (sozial-)psychologischen Forschung gängige Praxis ist, die Testpersonen, die an einem Experiment teilnehmen aus einer studentischen Population zu rekrutieren. Da sich diese Gruppe ja durchaus systematisch von anderen Gruppen unterscheiden könnte – so die Argumentation weiter – sind auch hier die Voraussetzungen für Signifikanztests keineswegs erfüllt.

Dieser Einwand trifft jedoch aufgrund der inhärenten Forschungslogik eines Experiments nicht zu. Angenommen man führt ein psycholinguistisches Lesezeitexperiment in einem Forschungslabor durch. In dem Experiment möchte man zum Beispiel testen, ob ein mit Hilfe sprachwissenschaftlicher Kriterien erstellter verständlichkeitsoptimierter fachsprachlicher Text (vgl. Wolfer, Hansen, & Konieczny, 2013) schneller gelesen werden kann als der ursprüngliche Text.

Eine Studentin betritt das Labor und bekundet ihre Teilnahmeabsicht. Entscheidend ist nun, dass die Versuchsleiterin per Zufall entscheidet, ob die betreffende Person in die Experimentalgruppe (optimierter Text) oder die Kontrollgruppe (Ausgangstext) kommt.

Der Randomisation an dieser Stelle kommt eine zentrale Bedeutung zu, da sie das wohl fundamentalste Problem jedweder empirischer Untersuchung löst (Angrist & Pischke, 2008, S. 9–18). Zeigt sich ein Unterschied zwischen den beiden experimentellen Gruppen hinsichtlich der Lesezeit, so ist dieser Unterschied nicht darauf zurückführbar, dass es sich um eine Studentin handelt und diese eben schneller liest - die Studentin hätte ja ebenso gut in die eine wie in die andere Gruppen eingeteilt werden können. Der Effekt muss deshalb darauf zurückgeführt werden, dass es sich um den Einfluss der experimentellen Manipulation handelt, weil die beiden Gruppen ja bis auf Zufallsschwankungen völlig identisch sind. So ist zum Beispiel nicht davon auszugehen, dass in der Kontrollgruppe nur die langsamen Leserinnen sind. Dies wäre zumindest bei einer halbwegs vernünftigen Stichprobengröße doch sehr unwahrscheinlich.

In der Korpuslinguistik könnte man als Lösung des Problems nun einfach dafür argumentieren, dass man sich mit Aussagen der folgenden Art begnügt: Es zeigt sich ein statistisch signifikanter Unterschied zwischen Korpus A und Korpus B. Dies hätte den Vorteil, dass man nicht davon ausgehen muss, dass die jeweiligen Korpora als Zufallsauswahl der jeweiligen sprachlichen Varietät fungieren und man nur beschreibt, was man in den jeweiligen Textsammlungen vorgefunden hat. In diesem Fall läuft jedoch ein statistischer Signifikanztest *per Definitionem* ins Leere. Zählt man zum Beispiel alle Instanzen eines bestimmten Modalverbs in Korpus A und vergleicht man diese mit dem Auftreten des gleichen Modalverbs in Korpus B, so hat man es überhaupt nicht mehr mit einer Stichprobe, sondern viel mehr einer Vollerhebung zu tun, weil man ja alle Elemente der jeweiligen Grundgesamtheit untersucht. Der inferente Schluss von der Stichprobe auf die Grundgesamtheit entfällt somit, weshalb sich Aussagen über die statistische Signifikanz eines Zusammenhangs eigentlich erübrigen.

Was folgt nun aus den hier angestellten Überlegungen?

Entscheidet man sich für eine rigorose Auslegung der Regeln der Inferenzstatistik, so gilt, dass statistische Signifikanztests eine mathematische Präzision vermitteln, die auf Grundlage von korpuslinguistischen Daten nicht haltbar ist. Aus diesem Grund müsste man strenggenommen generell auf die Verwendung von statistischen Signifikanztests verzichten. Andererseits gibt es auch gute Gründe (Diekmann, 2002, S. 600–601) die dafür sprechen die Berechnung von Signifikanztests zumindest als Orientierungshilfe für die Plausibilität eines Ergebnisses auch in korpuslinguistischen Untersuchungen beizubehalten.

Für welche der beiden Optionen man sich entscheidet, bleibt wohl letztlich Sache des individuellen Geschmacks. Ich hoffe jedoch in meinem Vortrag zu zeigen, dass Signifikanztests allein eine

sorgfältige inhaltliche Interpretation des erzielten Ergebnisses nicht ersetzen. Vielmehr sollten diese in jedem Fall durch Maße der Assoziations- bzw. Effektstärke ergänzt werden (Jann, 2005, S. 66–98), welchen unter Umständen sogar der Vorrang bei der Einordung des Forschungsergebnisses gegeben werden sollte.

Literatur

Angrist, J. D., & Pischke, J.-S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton NJ: Princeton University Press.

Baroni, M., & Evert, S. (2009). Statistical methods for corpus exploitation. In A. Lüdeling & M. Kytö (Hrsg.), *Corpus linguistics: An international handbook* (Bd. 2, S. 777–802). Berlin: De Gruyter Mouton.

Diekmann, A. (2002). *Empirische Sozialforschung: Grundlagen, Methoden, Anwendungen* (8. Aufl.). Reinbek: Rowohlt Taschenbuch Verlag.

Fahrmeir, L., Künstler, R., Pigeot, I., & Tutz, G. (2001). *Statistik: der Weg zur Datenanalyse ; mit 34 Tabellen*. Berlin [u.a.]: Springer.

Gries, S. T., & Berez, A. L. (noch nicht erschienen). Linguistic annotation in/for corpus linguistics. In N. Ide & J. Pustejovsky (Hrsg.), *Handbook of Linguistic Annotation*. Berlin, New York: Springer. Abgerufen von http://www.linguistics.ucsb.edu/faculty/stgries/research/InProgr_STG_ALB_LingAnnotCorpLing_HbOfLingAnnot.pdf

Jann, B. (2005). *Einführung in die Statistik*. München; Wien: Oldenbourg.

Lüdeling, A., & Evert, S. (2005). The emergence of productive non-medical -itis. Corpus Evidence and qualitative analysis. In S. Kepser & M. Reis (Hrsg.), *Linguistic Evidence. Empirical, Theoretical, and Computational Perspectives*. Berlin, New York: De Gruyter Mouton.

Schmid, H.-J. (2010). Does frequency in text instantiate entrenchment in the cognitive system? In D. Glynn & K. Fischer (Hrsg.), *Quantitative Methods in Cognitive Semantics: Corpus-Driven Approaches* (S. 101–133). Berlin, New York: de Gruyter.

Wolfer, S., Hansen, S., & Konieczny, L. (2013). *Are shorter sentences always easier? Discourse level processing consequences of reformulating texts.* Gehalten auf der European Society for Translation Studies 7, Gernershein. Abgerufen von http://www.fb06.uni-mainz.de/est/Dateien/EST_2013_abstract_booklet_web.pdf