

Gerd Simon

Ein Ähnlichkeitsmaß

Das Ähnlichkeitsmaß von dem hier die Rede sein soll, habe ich unter der Bezeichnung >Divergenz< („Unterschiedenheit“) erstmals in einem Vortrag („Textkritik und datenverarbeitende Maschine“) am 19. Dezember 1969 auf einer Veranstaltung des >Instituts für Kommunikation und Phonetik< (IPK) und des >Germanistischen Seminars< der Universität Bonn vorgestellt. Wie mir der Leiter des IPK, Gerold Ungeheuer, später mündlich mitteilte, wurde es an seinem Institut alsbald auf Spracherkennungsprobleme zugeschnitten. Ich selbst lieferte 1970 im Anhang meiner Dissertation¹ die mathematische Ableitung. Ein Verfahren, das ich für den Sonderfall der Kontamination entwickelte, stellte ich am 18. Juni 1971 in einem Vortrag auf einer Tagung im >Institut für Deutsche Sprache< (IDS) vor. Dieser Vortrag wurde 1978 – leicht überarbeitet – in einem Sammelband über die > Maschinelle Verarbeitung altdeutscher Texte< abgedruckt.²

Mir war damals und ist im übrigen auch heute nicht wichtig, ob das grundlegende Maß der Ähnlichkeit bzw. der Divergenz schon zuvor von anderen entdeckt bzw. entwickelt wurde. Forschungen kommen nicht selten unabhängig zum gleichen Ergebnis. Der Erstentdeckungsanspruch ist nur da von Bedeutung, wo jemand sich dafür gepriesen oder ausgezeichnet wissen will – was mir nur da relevant gewesen wäre, wo eine Aussicht auf Forschungsgelder für zukünftige Studien bestand – oder wo ein Plagiatsverdacht besteht. In diesen Verdacht hat mich bisher niemand gebracht und andere in diesen zu bringen, hätte mir nur Zeitverlust bedeutet. Das Maß habe ich in einer einfachen mathematischen Formel zusammengefasst:

$$\delta = 1 - (1 - L/N)^{1/i}; \quad i \geq 2$$

Dabei bedeutet δ die Divergenz oder Ähnlichkeit von zwei und mehr Gesamtheiten, L die Anzahl der verschiedenen Elemente in diesen Gesamtheiten, N die Anzahl aller Elemente in diesen und i die Zahl der Gesamtheiten.

¹ Die erste deutsche Fastnachtsspieltradition. Zur Überlieferung, Textkritik und Chronologie der Nürnberger Fastnachtsspiele des 15. Jahrhunderts (mit kurzen Einführungen in Verfahren der quantitativen Linguistik). Lübeck, Hamburg: Matthiesen. 1970
(= Germanistische Studien 240)

² Simon, Gerd: Zur Theorie der Kontamination: in: Maschinelle Verarbeitung altdeutscher Texte. Beiträge zum Symposium Mannheim 11./12. Juni 1971 (Hg. v. W. LENDERS und H. MOSER). Berlin: Erich Schmidt. 1978, S. 108-116

Ich habe diese Formel seinerzeit in einem relativ harmlosen Forschungsgebiet der Philologie, genauer der Textkritik, angewandt und damit versucht, die Qualität der Überlieferung eines Textes zu bestimmen. Wenn ich mich nicht täusche, wurde sie in der Textkritik nie aufgegriffen. Unter den gut zwei Dutzend Rezensionen, die meine Dissertation erfuhr, hielt nur eine einzige das Niveau meiner Ausflüge in die Mathematik für erwähnenswert und ihr Verfasser war nach Auskunft des Herausgebers der Zeitschrift, in der die Rezension erschien, ein amerikanischer Mathematiker. Ich habe diesen übrigens seinerzeit angeschrieben, aber nie einen Antwort erhalten.

Da es sich um ein allgemeines Vergleichsmaß handelt und ich schon in meinem 1956 begonnenen Philosophiestudium lernte, dass wissenschaftliche Forschung allem voran Vergleichen hieß, also Gemeinsamkeiten, Ähnlichkeiten und Unterschiede feststellen und möglichst Exaktizieren, war mir auch sofort klar, dass sich die Formel sogar einfacher als in den Philologien in anderen Disziplinen anwenden ließ, zumindest in solchen, die so weit gediehen waren, dass in ihnen ein System von klar voneinander verschiedenen Elementen ausgemacht war.

Dass das sogar in den Sprachwissenschaften nicht so einfach war, vermittelte mir mein Studium vor allem des linguistischen Strukturalismus. Das sei hier wenigstens andeutungsweise ausgeführt. Es gibt Bücher, die ein und denselben Text in Hunderten von Sprachen wiedergeben.¹ Will man die Verwandtschaft zwischen diesen Sprachen auf Grund dieses Textes ermitteln, kann man selbst bei alphabetischen Schriften nicht einfach die Buchstaben als Element nehmen. Denn sie entsprechen in keiner Sprache – am ehesten noch in Esperanto – exakt den gesprochenen Lauten. Wie stark die Schrift von den Sprechlauten abweichen kann, erschließt sich jedem sofort, der Englisch lernt. Die Lautung ist hier von der Schrift aus nur schlecht vorhersagbar und muss gesondert erlernt werden. Eine nicht geringe Rolle spielen hier auch Morphologie, Syntax und Lexikologie.

Schon im 19. Jahrhundert bemühte man sich um eine Lautschrift.² Vereinfacht findet man sie noch in den Fremdsprachenlexika im Anschluss an ein Lexem in Klammern hinzugefügt. Die Lautschrift setzt aber eine umfassende Analyse der Laute voraus. Da gab es das Riesenproblem, dass mit physikalischen Methoden, v.a. den Visible-Speech-Geräten (am bekanntesten wurden die Sonographen) nur Lautströme ausgemacht werden konnten, keine einzelnen Lau-

¹ Britische und Ausländische Bibelgesellschaft: Gottes Wort in vielen Sprachen. Proben von 543 Sprachen... London. 1921.

² Nach vielen Vorschlägen u.a. von dem Tübinger Linguisten Moritz Rapp kam es durch die 1886 gegründete >International Phonetic Association< zu einer ersten Kodifizierung, die im Kern noch heute weltweit gilt.

te. Hinzu kam, dass kein Sprecher willentlich den gleichen Laut mit der gleichen Frequenz wieder so traf, wie er ihn gerade erst ausgesprochen hatte.

Es waren die Strukturalisten, die angesichts dieser höchst unübersichtlichen Situation einen Weg fanden, zu objektiven Erkenntnissen zu kommen.¹ Sie ermittelten an der Lautung distinctive features, Merkmale, die nach experimenteller Änderung zu sinnlosen oder zu Bedeutungsunterschieden führten, und umfassende Systeme dieser distinctive features, nach denen die Laute als Kombination einer begrenzten Anzahl dieser Merkmale aufgefasst werden konnten.

Um die Verwandtschaft von Sprachen ausfindig und messbar zu machen, war es der Königsweg, diese distinktiven Merkmale als Elemente zugrunde zu legen. Dabei konnten ältere Unterscheidungen, z.B. nach dem Artikulationsort (Lippen, Zähne, Rachen usw.) und der Artikulationsart (stimmhaft oder stimmlos, Explosiv- oder Reibelaute usw.) aufgegriffen und mit akustischen Merkmalen zu einem Elementesystem (Merkmalsbündeln) kombiniert werden, das auf jede Sprache anwendbar war.

Dabei ging man zumeist folgendermaßen vor:

- Auswahl einfacher verbreiteter Wörter. Ein Anfangsverdacht auf Grund ihrer Bedeutung ist nicht notwendig, aber durchaus hilfreich. Auch die Beschränkung auf KVK-Wörter² erweist sich als keineswegs dumm.
- Wörter, deren Laute nur in einem Merkmal systematisch, wenn nicht ausnahmslos abweichen, stehen in größerem Verwandtschaftsverdacht als gleichlautende, da letztere auf Entlehnungen beruhen können. D.h. Ähnlichkeit liefert hier stärkere Hinweise als Identität.
- Einzelbeispiele sind auch da nur von geringer Bedeutung, wo sie offensichtlich nicht auf Entlehnung beruhen. (Beispiel mam – ‚Brust‘, ‚Mutter‘). Will man sicher gehen, muss es ein Fülle gleichgearteter Beispiele geben. Die Suche nach „ausnahmslosen Gesetzen“ hat sich umgekehrt als unproduktiv erwiesen.

Da gerade eng verwandte Sprachen wie das Hoch- und das Niederdeutsche sich durch das unterscheiden, was man seit dem 19. Jahrhundert Lautverschiebung nannte, was faktisch aber nur den Übergang innerhalb ähnlicher Merkmalsbündel betrifft (z.B. $\underline{d} > t$; $\underline{e} > \bar{i}$; $\underline{p} > \underline{f}$ – d.h.

¹ Hier wäre v.a. einer der wenigen deutschen Strukturalisten zu nennen: Eberhard Zwirner. s. dazu <http://homepages.uni-tuebingen.de/gerd.simon/strukturalismus1.htm>

² KVK-Wörter sind Wörter mit der Struktur: Konsonant-Vokal-Konsonant

ndd. dēp > nhd. tief) war es für die Grobermittlung von Verwandtschaftsbeziehungen naheliegend, ähnliche distinktive Merkmale zu Klassen zusammen zu fassen, wie das – konzentriert auf die wichtigsten Lautgruppen – folgende Vergleichsanweisung tut:

	[Anlaut]	Inlaut	[Auslaut]
KVK à	$\left\{ \begin{array}{c} \text{KL} \\ \text{KD} \\ \text{KG} \\ \text{KN} \end{array} \right\}$	$\left\{ \begin{array}{c} \text{VLH} \\ \text{VLD} \\ \text{VKH} \\ \text{VKD} \end{array} \right\}$	$\left\{ \begin{array}{c} \text{KL} \\ \text{KD} \\ \text{KG} \\ \text{KN} \end{array} \right\}$

Dabei gilt:

KVK à Konsonant-Vokal-Konsonant

KL à Konsonant, Labial (z.B. p, b, f)

KD à Konsonant, Dental (z.B. t, d)

KG à Konsonant, Guttural (z.B. g, k)

KN à Konsonant, Nasal¹ (n, m)

VLH à Vokal, lang, hell (z.B. ī, ē)²

VLD à Vokal, lang, dunkel (z.B. ā, ō, ū)

VKH à Vokal, kurz, hell (z.B. ī, ē)

VKD à Vokal, kurz, dunkel (z.B. ǣ, ǒ, ǔ)

Danach wären nd. dēp und nhd. tief der Struktur KVK -> [KD] [VLH] [KL] zuzuweisen. Will man über diese Grobanalyse hinaus den Grad der Ähnlichkeit bestimmen, müsste man die Einzelmerkmale der Merkmalsbündel genauer bestimmen. Dann würde sich in diesem Fall zeigen, dass ndd. dēp größere Ähnlichkeit zu engl. deep aufweist, das dīp gesprochen wird. Dieses Einzelbeispiel sollte aber nicht vorschnell zu der These verführen, dass das Nieder-

¹ In manchen Sprachen sind die Übergänge zwischen Konsonant und Vokal fließend. Das betrifft v.a. nasale (n, m) und intermittierende (l, r) Laute. Man spricht dann von Halb-Vokal oder auch –konsonant.

² In den meisten Sprachen sind dazu auch die Diphtonge (ei, oi, au) zu rechnen. Sie verhalten sich jedenfalls wie diese.

deutsche mit dem Englischen verwandter ist als mit dem Hochdeutschen. Dazu im Folgenden einige Hinweise.

Studien zur Geschichte von Sprachen beiseite zu lassen, mag für die systematische Analyse erstrebenswert erscheinen, verzichtet aber unnötigerweise auf Möglichkeiten der Interpretation und Erklärung. Vorinformationen wie die, dass das Englische eine Mischsprache auf Grund zahlreicher Sprachkontakte ist, wobei (kaum noch nachweisbare) Substrate durch Superstrate wie das Keltische, Lateinische, Angelsächsische und Französische seine gegenwärtige Gestalt erhielten, sollten auch bei massenhaften Belegen zu Ergebniseinschränkungen führen, z.B. dazu, dass das Niederdeutsche nur mit dem angelsächsischen Anteil am Englischen näher verwandt ist als das Hochdeutsche. Wenn die Synonyma mit in die systematische Analyse einbezogen werden, was methodisch sogar unabdingbar, wenn auch nicht leicht umzusetzen wäre, dürften auch solche Ergebnisse nicht allzu sehr abweichen. Die systematische Analyse allein könnte die plötzlichen Unterschiede bei sonstiger Ähnlichkeit aber nicht erklären.

Grundsätzlich gilt außerdem: Sprachliche Phänomene verhalten sich nicht so simpel wie etwa Tannennadeln. Sie sind Ordnungsgebilde und darum nur selten und dann auf lautlicher Ebenen zufallsverteilt, d.h. Mittelwerte und Streuungsmaße sagen hier nichts Signifikantes über sie aus¹; weiterverarbeitet führen sie sogar zu eindeutig falschen Interpretationen. Aus diesen Gründen ist es sinnvoll, von vornherein von einer weitaus höheren Grundgesamtheit mit nicht allzu starker Abweichung in der Zahl aller Elemente auszugehen, als in den Naturwissenschaften üblich.

Die von mir vorgeschlagene Grobanalyse würde die obige Ähnlichkeitsformel natürlich nicht tangieren. Ich habe nur das Beispiel der Ähnlichkeitsbeziehungen zwischen Sprachen gewählt, um deutlich zu machen, wie differenziert diese Formel in einzelnen Fällen angewendet werden muss. Immerhin basiert diese Grobanalyse auf 64 Dreilautwörtern und gestattet damit einen relativ zuverlässigen Vergleich, zuverlässig in Bezug auf das Faktum der Verwandtschaft, nur sehr grob in Bezug auf den Grad der Verwandtschaft.

Die Ähnlichkeitsformel lässt sich – wie erwähnt – grundsätzlich auf alle Wissenschaften anwenden, die über ein exaktes Elementesystem verfügen. Hier möchte ich nur noch kurz auf eine Anwendung in der Genetik hinweisen.

¹ Das hatte schon der Emigrant Gustav Herdan (Language as Choice and Chance. Groningen 1956) festgestellt, glaubte diese Werte freilich durch andere Werte ersetzen zu können. Diese führten aber wohl bestenfalls bei der von ihm untersuchten Textpopulation zu zuverlässigen Ergebnissen.

Heute werden Ergebnisse der Genforschung in der Kriminalistik angewandt. Die vor mehr als 40 Jahren vorgestellte und oben wiedergegebene Ähnlichkeitsformel kann z.B. die DNA-Muster in einer Zahl ausdrücken. Ich habe das nie getan, weil ich kein Genforscher bin, halte das aber für weitaus leichter praktikierbar als beim Sprachenvergleich. Allerdings in dem heute wieder für möglich gehaltenen Fall, dass Lamarck gegenüber Darwin recht hat, dass also bei der Rekonstruktion der Deszendenz mit Kontaminationen zu rechnen ist, wird das auch hier komplizierter. Auch für diesen Fall entwickelte ich ein Verfahren.¹ Auch in der Kriminalistik spielt dieser Fall eine vergleichbare Rolle, wenn es z.B. um Nachfahren einer Geschwister- oder einer anderen Verwandtschaftsbeziehung geht.

Hier ergeben sich freilich auch noch ganz andere, gewichtigere Probleme. Ich habe mir von Anfang an die Frage gestellt: was würde ein Rassist wie Heinrich Himmler mit dieser Formel gemacht haben? Ich selbst habe in einem Archiv die Reaktion Himmlers auf eine [Falsch-] Meldung in einer amerikanischen Fachzeitschrift gefunden und später veröffentlicht, wonach Experimente mit dem Herbstzeitlosengift Colchizin zu Genveränderungen führten.² Himmler hätte das gnadenlos für seine Menschheitsverbrechen eingesetzt. Das hätte er freilich auch mit der DNA-Analyse selbst getan, die ja mit dem Ähnlichkeitsmaß nur auf einen Vergleichswert gebracht wird. Ich will aber nicht verhehlen, dass Bedenken berechtigt waren, die obige Formel zu verbreiten.

An sich pflege ich in wissenschaftlichen Publikationen keine Aufrufe zu inkorporieren. Hier mache ich aber wegen dieser Bedenken eine Ausnahme, obwohl der folgende Aufruf eigentlich am Schluss aller, vor allem meiner 3.-Reich-Forschungen stehen könnte:

Der Rassismus ist keineswegs tot. Es gilt, ihm mit argumentativen Mitteln, die jedenfalls seine Gewaltaktionen nicht einfach kopieren, zu verhindern, dass er irgendwo auf der Welt wieder Einfluss gewinnt. Man rechne damit, dass er unter dem Mantel des Antirassismus die Exekutive dazu bringt, immer rassistischere Aktionen und Gesetze zu praktizieren. Man rechne aber auch damit, dass er umgekehrt bei allem Aktionismus durch Zerreden und böartige Kritik im Vorfeld von Aktivitäten erst einmal alles Antirassistische zu verhindern sucht. Man setze eher auf zäh wiederholte Nadelstiche, statt auf spektakuläre Brutalitäten, auf radikale Demokratie statt auf Hierarchisierung und auf nüchterne Vermittlung von Fakten statt auf wild gestikulierende Propaganda. Gegen den Geist der Sklavenhaltung, der vor mehr als 10 000 Jahren in die Mensch-

¹ Simon, Gerd: Zur Theorie der Kontamination. in: Maschinelle Verarbeitung altdeutscher Texte. Beiträge zum Symposium Mannheim 11./12. Juni 1971 (Hg. v. W. LENDERS u.a.). Berlin: Erich Schmidt. 1978, S. 108-116.

² <http://homepages.uni-tuebingen.de/gerd.simon/genetik1.htm>

heit fuhr, der unzählige Varianten hervorbrachte (und nicht nur den Rassismus), hilft nur ein langer Atem, kein Alles-oder-Nichts und schon gar nicht subito. Man lasse die Finger von den Mitteln der Sklavenhaltung von den Kritikverböten über die Verführung durch Werbung und Einschüchterung bis zur Folter und Hinrichtung. Das führt bestenfalls dazu, dass die als rassistisch entlarvte Fratze ihre Maske wechselt.