IDS LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

PRESSEMITTEILUNG

17.10.2023

Studie mit über 2000 Sprachen zeigt: Nicht alle Sprachen sind gleich komplex

Die moderne Sprachwissenschaft nahm lange an, dass die ca. 7000 Sprachen auf der Welt mehr oder weniger gleich komplex sind. Eine neue Studie des Leibniz-Instituts für Deutsche Sprache in Mannheim deutet nun darauf hin, dass sich Sprachen durchaus in ihrer Komplexität unterscheiden – Sprachen, die von mehr Menschen gesprochen werden, erreichen dabei höhere Komplexitätswerte. Diese Untersuchung ist die größte jemals durchgeführte korpuslinguistische sprachvergleichende Studie: Sie umfasst über 6500 Dokumente mit insgesamt mehr als 3,5 Milliarden Wörtern in über 2000 Sprachen, die von mehr als 90% der Weltbevölkerung als Muttersprache gesprochen werden.

Lange wurde in der modernen Linguistik davon ausgegangen, dass alle Sprachen mehr oder weniger gleich komplex sind. Die Grundidee dieser sogenannten „equi-complexity hypothesis“ (Hypothese der gleichen Komplexität) ist, dass ein „Ausgleich“ zwischen verschiedenen sprachlichen Ebenen stattfindet. Eine Sprache, bei der zum Beispiel die Stellung der Wörter im Satz sehr beschränkt ist (die auf der Satzebene also wenig komplex ist), kann das über reichhaltigere Wortbildungsmittel (wie z. B. Vor- und Nachsilben) ausgleichen. Betrachtet man sehr viele Sprachen, müsste das dazu führen, dass sich die allgemeine Komplexität einer beliebigen einzelnen Sprache nicht sonderlich stark von jener anderer Sprachen unterscheidet.

Möchte man dies anhand eines Vergleichs sehr vieler Sprachen überprüfen, muss man diese „allgemeine Komplexität“ einer Sprache irgendwie messen – und es war bisher nicht klar, wie das möglich sein sollte. Alexander Koplenig, Sascha Wolfer und Peter Meyer vom Leibniz-Institut für Deutsche Sprache (IDS) gehen diese Frage in ihrer nun in Nature Scientific Reports veröffentlichen Studie aus informationstheoretischer Perspektive an. Anstatt die Komplexität jeder sprachlichen Ebene (Wort, Satz und Text) einzeln zu messen (wobei völlig unklar wäre, wie diese Ebenen später kombiniert werden sollen), gehen sie davon aus, dass sprachliche Komplexität über eine Art „Sprachspiel“ gemessen werden kann.

Wenn man Menschen einen Satzbeginn vorlegt wie *Auf dem Gipfel des Berges angekommen, genossen sie den atemberaubenden …* werden die allermeisten eine gute Vorstellung davon haben, wie dieser Satz weitergehen könnte. *Sonnenuntergang* wäre eine hervorragende Fortsetzung, *Rotwein* ginge auch, ist aber weniger wahrscheinlich. Andere Fortsetzungen sind sehr unwahrscheinlich oder gar ausgeschlossen – zum Beispiel ein weiterer Artikel wie *die* oder ein Nomen mit femininem Genus wie *Aussicht*. Das funktioniert nicht nur mit Wörtern, sondern auch mit Buchstaben. Welcher Buchstabe folgt nach *Die Frau sag*? Höchstwahrscheinlich doch ein *t*. Interessanterweise schnitten Menschen hier bis vor gar nicht langer Zeit deutlich besser ab als Maschinen. Der Grund: Menschen sind ihr ganzes Leben lang Sprache ausgesetzt. Ob gesprochen oder geschrieben, ob im öffentlichen oder im privaten Raum, die ganze Zeit trainieren und nutzen wir die internen Gesetzmäßigkeiten der Sprache – und deshalb können wir in den allermeisten Fällen einigermaßen gut vorhersagen, welches Wort oder welcher Buchstabe als nächstes kommt. Oder wir wissen zumindest, wann wir es nicht können (zum Beispiel ganz am Anfang eines Satzes oder Gesprächs).

Genau diese Idee machten sich die Forscher des IDS nun zu Nutze. Gibt man einem Computer große Mengen an Textmaterial, kann dieser die statistischen Gesetzmäßigkeiten von Sprachen „lernen“, um dieses Vorhersagespiel möglichst gut spielen zu können. Wie gut die Maschine dabei am Ende des Lernprozesses abschneidet, ist ein Maß dafür, wie komplex das sprachliche Material war, das dem Computer vorgelegt wurde – man nennt dieses Maß die „Entropierate“. Möchte man nun die Komplexität vieler Sprachen miteinander vergleichen, muss man darauf achten, dass der vermittelte Inhalt möglichst gleich ist. Dafür haben die Forscher 41 mehrsprachige Textkollektionen untersucht, so zum Beispiel Übersetzungen der Bibel oder des Korans in viele Sprachen. In jeder dieser Kollektionen

kann man davon ausgehen, dass über alle verfügbaren Sprachen hinweg ungefähr der gleiche oder doch – wie im Fall von Filmuntertiteln oder verschiedenen Sprachversionen eines Computer-Betriebssystems – wenigstens grundsätzlich vergleichbarer Inhalt vermittelt wird. Insgesamt enthalten die Textkollektionen über 6500 Dokumente in über 2000 Sprachen, die von über 90% der Weltbevölkerung gesprochen werden.

Wenn man nun davon ausgeht, dass alle Sprachen mehr oder weniger gleich gut vom Computer vorhersagbar sind – wie die oben beschriebene Hypothese der gleichen Komplexität besagt – sollten sich über die Sprachen der Welt hinweg keine systematischen Unterschiede finden: Ist eine Sprache in einer Kollektion komplexer als eine andere, sollte das in einer anderen Kollektion wieder andersherum sein – am Ende sollte sich alles ausgleichen. Doch das ist nicht der Fall. In der IDS-Studie zeigte sich vielmehr ein systematisches Bild: Sprachen, die in einem Dokument schwieriger vorherzusagen sind als andere, sind auch in vielen anderen Dokumenten komplexer. Und es zeigte sich noch eine andere Systematik: Sprachen, die von mehr Menschen als Muttersprache gesprochen werden, sind komplexer als Sprachen, die von weniger Menschen gesprochen werden. Das gilt nicht nur innerhalb sehr „großer“ Sprachen wie Mandarin-Chinesisch, Englisch oder Deutsch, sondern ist über alle untersuchten Sprachen verallgemeinerbar. Auch ist der Effekt nicht beschränkt auf bestimmte Sprachfamilien oder Zeichensysteme.

Woher dieser Effekt stammt, ist bisher noch weitestgehend ungeklärt, aber erste Beobachtungen zeigen, dass Sprachen mit größeren Sprachgemeinschaften nicht nur komplexer sind, sondern auch kürzere Texte „produzieren“. Diese kürzeren Texte (bei gleichbleibendem Inhalt) könnten vorteilhaft im gesellschaftlichen, ökonomischen und technologischen Zusammenleben sein, weil sie einfacher zu übermitteln und zu speichern sind. Weitere Studien zu diesem Thema werden momentan am IDS erarbeitet.

Originalarbeit: Koplenig, Alexander & Wolfer, Sascha & Meyer, Peter (2023). A large quantitative analysis of written language challenges the idea that all languages are equally complex. Scientific Reports 13(1). 15351. DOI: https://doi.org/10.1038/s41598-023-42327-3

Ansprechpartner:

Dr. Sascha Wolfer
Leibniz-Institut für Deutsche Sprache
R 5, 6-13
D - 68161 Mannheim
Tel.: +49 621 / 1581 - 439
E-Mail: wolfer@ids-mannheim.de

Pressekontakt:

Dr. Annette Trabold
Leiterin Öffentlichkeitsarbeit
Leibniz-Institut für Deutsche Sprache
R 5, 6-13
D - 68161 Mannheim
Tel.: + 49 621/ 1581-119
E-Mail: trabold@ids-mannheim.de