COMPUTER

Hindi in vier Wochen

Ein Deutscher macht in Kalifornien Furore mit neuartigen Übersetzungsprogrammen. In Rekordzeit bringt er dem Rechner fremde Sprachen bei – obwohl er ihrer gar nicht mächtig ist.

VOLKER CORELL

Übersetzungspionier Och: *Sprache mit roher Gewalt enträtselt*

Wie ein Sprachgenie wirkt Franz Josef Och nicht gerade. Sein Hochdeutsch klingt fränkisch, sein Englisch deutsch, und sein Italienisch reicht mal eben für die Trattoria. „Abitur“, gesteht er, „habe ich auch nicht.“

Trotzdem hat sich Och, 31, auf den Radarschirm des Pentagon und der Geheimdienste katapultiert – als Sprachbezwinger. Vor einem Jahr hat ihn die University of Southern California von der TH Aachen nach Los Angeles gelockt. Dort sitzt er hoch oben in einem schwarzweißen Büroturm mit kalifornischem Panoramablick – und an seinem Rechner vollbringt er das Mirakel, das ihn für Verbrecher- und Terroristenjäger so interessant macht.

Och kann seinem Computer verblüffend schnell beibringen, Texte in jeder beliebigen Sprache selbständig ins Englische zu übersetzen. Ist ein Programm einmal fertig, braucht es für eine DIN-A4-Seite nicht länger als eine Minute.

Finanziert wird Ochs Arbeit zum größten Teil von der Darpa, der Forschungsorganisation des Pentagon. Hält er einen Vortrag, dann gesellen sich gern auch Analysten des Abhör- und Spionagedienstes NSA (National Security Agency) hinzu. Auf einen wie Och haben sie gewartet, denn in Zeiten des Terrors versuchen die Amerikaner mit Riesenaufwand, Herr zu werden über die babylonische Sprachverwirrung auf Erden.

22 Millionen Dollar gibt das Pentagon allein in diesem Jahr aus für das ehrgeizige „Tides“-Projekt, das die weltweite Suche nach Feinden und Bedrohungen revolutionieren soll: Telefonate in fremden Sprachen zum Beispiel soll es abhören, in Text verwandeln, auf Englisch übersetzen und zusammenfassen. Die US-Strategen interessieren sich für alles: für arabische Zeitungen, chinesische E-Mails, Radiosendungen auf Urdu, Chatrooms in Bengali. Die Sprachbarrieren zu durchbrechen gilt den Amerikanern nunmehr als Gebot der nationalen Sicherheit.

FALK HELLER / ARGUM

NSA-Abhöranlage (in Bad Aibling): *Geld vom Pentagon*

Kürzlich hat Och teilgenommen an einem Wettbewerb der Darpa. Vier Teams von Maschinenübersetzern sollten gegeneinander antreten. Niemand wusste, welche Sprache ihnen zu knacken aufgegeben werden würde, deshalb war der Nervenkitzel groß, als am 1. Juni um 22.55 Uhr die E-Mail einschlug: „Die Überraschungssprache ist Hindi … Viel Glück!“

Vier Wochen später war Och fertig. Sein Computer konnte das hoch komplizierte Hindi übersetzen – bei weitem nicht gut genug für Literatur, aber gerade ausreichend, damit sich ein Amerikaner mit etwas Geduld den Inhalt einer Zeitung erschließen kann. Ochs Programm überzeugte die Jury: In allen Kategorien ließ er seine Konkurrenten hinter sich.

Bis März dieses Jahres hatte Och alle EU-Amtssprachen ins Englische übertragen, sogar das verteufelt schwere Finnisch, dazu noch Chinesisch, Japanisch und Arabisch, insgesamt 15 Übersetzungen.

Och schafft all das ohne Grammatik- und Vokabelheft, und außerdem jongliert er mit den Sprachen so schnell wie sonst nur das Personal von Science-Fiction-Filmen: Für eine Grobversion von Cebuano, eine Sprache, die von 19 Millionen Menschen auf den Philippinen gesprochen wird, brauchte er nicht mehr als zehn Tage. „Notfalls“, prahlt er, „reichen auch Stunden.“

Ist Och ein Wunderkind? Ein Scharlatan? Ein Genie? Nichts dergleichen: Och ist Informatiker.

Nach der Realschule hat er sich über die Fachoberschule durchgearbeitet bis zum Universitätsstudium, wo er begann, sich für die so genannte Maschinenübersetzung zu interessieren. Das Feld galt als nicht besonders aussichtsreich: Seit den fünfziger Jahren wollten die Amerikaner Automaten haben, um russische Texte ins Englische zu übertragen. Aber trotz Jahren teurer Forschung hörten die Maschinen nicht auf, Unsinn zu brabbeln. Ende der sechziger Jahre gaben die meisten Entwickler schließlich auf: Sprache schien für Geräte schlicht zu kompliziert.

Revidiert wurde dieses Urteil kürzlich erst. „In letzter Zeit sind Riesenfortschritte gelungen“, sagt Och. Hochschulen und Firmen haben das Fach wiederbelebt, einfache Übersetzungsprogramme stehen bereits in den Geschäften.

Zwar liest sich, was diese Programme an Stummeltexten produzieren, oft wie die berüchtigten Betriebsanleitungen aus Südkorea. Doch das ändert sich – dank immer schnellerer Computer und einer neuen Denke: Viele Maschinenübersetzer versuchen nicht mehr, möglichst alle Regeln einer Sprache aufzuschreiben und für den Rechner anwendbar zu machen. Sie gehen nun ei-

nen Weg, den Forscher des Computerriesen IBM Anfang der neunziger Jahre ausgekundschaftet haben: Ob Albanisch oder Zulu – sie enträtseln eine Sprache einfach mit der rohen Gewalt von Statistik und Rechenkraft.

„Wir sagen dem Computer nicht, wie er übersetzen soll“, erklärt Och. „Wir lassen ihn das einfach selbst lernen.“ Und wenn er dann die Methode der „statistischen Übersetzung“ erklärt, geraten seine Zuhörer ins Staunen – und ins Grübeln, ob er sie nicht verschaukelt. Denn als Erstes nimmt der Franke die Bibel zur Hand: Am Anfang war das Wort.

Die Bibel ist in jede bedeutsame Sprache der Welt übersetzt und zudem leicht verfügbar. Deswegen liefert sie Och meist den ersten „Paralleltext“ – seine wichtigste Ressource. Er füttert seinen Rechner mit der Bibel auf Englisch und der Bibel auf Arabisch, außerdem sucht er möglichst viele weitere Beispieltexte in menschengemachter Übersetzung: Meldungen von internationalen Nachrichtenagenturen, Handbücher, amtliche Transkriptionen der Uno.

Wenn der Computer zum Beispiel weiß, dass das arabische *radschul kabir* zu Deutsch „großer Mann“ bedeutet und *radschul samin* für „fetter Mann“ steht, dann erschließt das Elektronenhirn, dass *radschul* „Mann“ heißt, *kabir* „groß“ und *samin* „fett“.

Paralleltext ist so etwas wie der Stein von Rosette des Digitalzeitalters: Die berühmte Tafel aus der Zeit um 200 vor Christus, 1799 in Ägypten gefunden, trägt dieselbe Inschrift in Griechisch, Demotisch und in Hieroglyphen. Mit ihrer Hilfe konnte der Linguist Jean François Champollion 1822 das Geheimnis der Hieroglyphen lüften – nach 13 Jahren Arbeit.

Je größer Ochs Fundus an Paralleltext, desto raffinierter die Übersetzung. Für Arabisch bediente er sich aus einer Textsammlung von 150 Millionen Wörtern, was in etwa der Textmenge von 30 SPIEGEL-Jahrgängen entspricht. Für den Schnellschuss Hindi musste sich Och mit rund 3 Millionen Wörtern begnügen, immerhin noch etwa 30-mal so viele, wie in dieser Ausgabe stehen. Gut gefüttert, kann der Rechner sogar idiomatische Ausdrücke erkennen und richtig interpretieren: Wenn ein Holländer zum Beispiel zugibt, „vom Geschäft keinen Käse gegessen zu haben“, dann bedeutet das schlicht, er habe vom Geschäft keine Ahnung.

Aber solche Dinge interessieren Och nur am Rande. „Ich selbst muss nicht viel wissen von den Sprachen“, sagt er. Seine eigentliche Leistung besteht aus den rund 10 000 Zeilen Computercode, in denen festgeschrieben steht, nach welchen statistischen Modellen sich der Rechner im Meer von zweisprachigem Beispieltext orientiert. Dass er damit bessere Ergebnisse erzielt als die Programme rechtschaffener Linguisten, macht ihn bei denen nicht gerade beliebt.

Von Perfektion ist auch Och natürlich weit entfernt. Nur manche seiner Übersetzungen klingen rund, viele hingegen weltfremd: Der Satzbau ist oft falsch, die Stilebenen gehen durcheinander, Deklinationen und Tempora scheinen mitunter Glückssache zu sein. „Die Texte“, räumt er ein, „sind unschön zu lesen.“ Niemals werde ein Computer ein Gedicht übersetzen.

Aber Ochs Geldgeber schert das wenig. Militärs und Geheimdienste interessieren sich nur für Antworten auf die vier W: Wer? Was? Wann? Wo? Gemäß der aktuellen Weltlage interessieren sie sich vor allem für eine Sprache: Arabisch.

ARCHIVO ICONOGRAFICO / COR

Stein von Rosette (im British Museum in London)
Geheimnis der Hieroglyphen gelüftet

Weil die Forscher dafür am leichtesten Gelder bekommen, ist Arabisch zurzeit die Sprache, die von Rechnern am besten beherrscht wird. Die Wunschliste des Pentagon lässt erahnen, auf welche Regionen die Strategen ihr Augenmerk richten: Neuerdings arbeiten Ochs Institutskollegen auch an Dari, gesprochen in Afghanistan, und Farsi, gesprochen in Iran.

Kürzlich haben die Tester der Normierungsbehörde National Institute of Standards and Technology nach den besten Übersetzungsprogrammen für Arabisch und Chinesisch gesucht. Im Rennen waren 7 kommerzielle Produkte und 16 aus den Denkstuben der internationalen Forscher.

Gewonnen hat jedes Mal der Mann ohne Abitur aus Pretzfeld, der sich nun in Kalifornien verdingen muss.

In Europa sah Och für sich keine Zukunft: „Gerade die Europäer mit ihrem Sprachengewirr sollten sich um Maschinenübersetzung kümmern“, sagt er. „Aber vergleichbare Projekte gibt’s da nicht.“

Marco Evers