ALLE ZWEI WOCHEN VERSCHWINDET EINE SPRACHE

Thomas Widlok

Elf von einigen Tausend verbliebenen Sprecherinnen und Sprechern der Klicklautsprache ≠Akhoe Hai//om im nördlichen Namibia

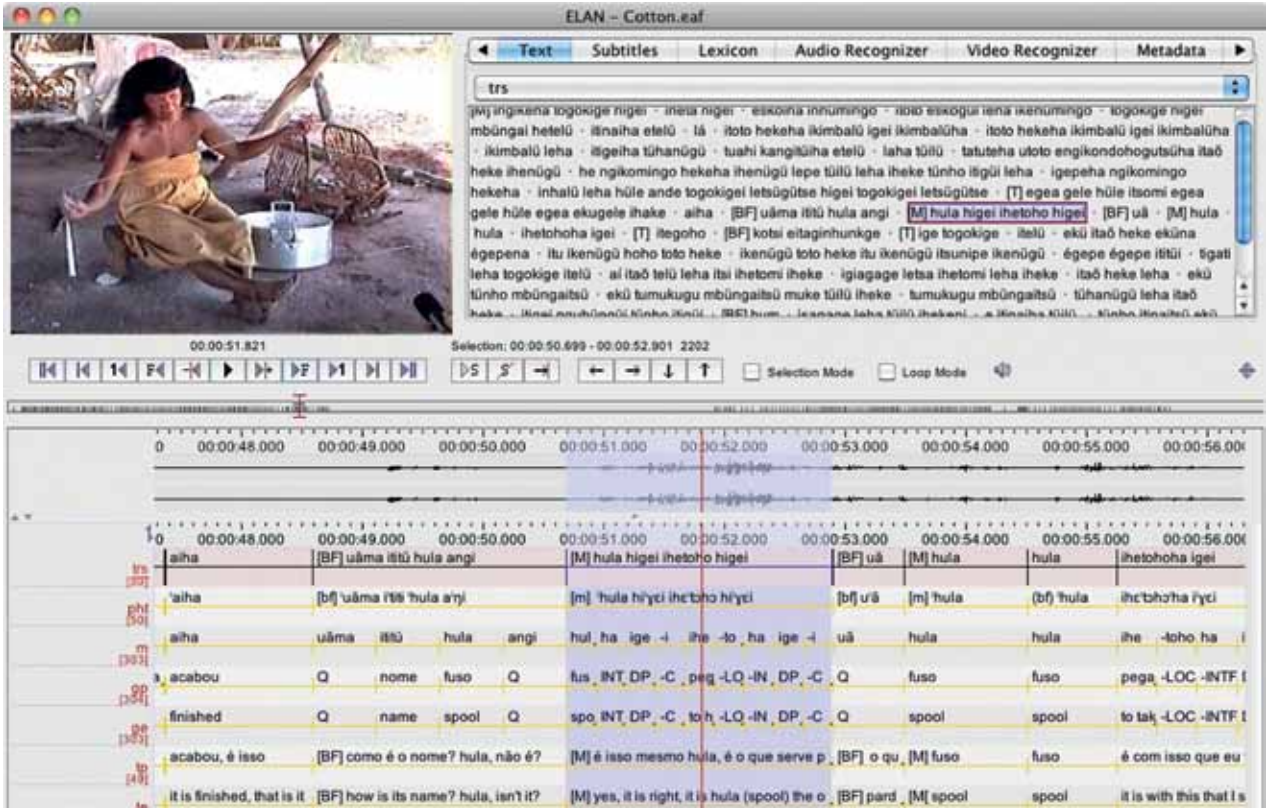
DIE AKADEMIE DOKUMENTIERT IN DEM PROJEKT »THE LANGUAGE ARCHIVE« (TLA) GEMEINSAM MIT DER KÖNIGLICH-NIEDERLÄNDISCHEN AKADEMIE DER WISSENSCHAFTEN UND DER MAX-PLANCK-GESELLSCHAFT VOM AUSSTERBEN BEDROHTE SPRACHEN

Von Wolfgang Klein

Es gibt derzeit schätzungsweise 6.000 bis 7.000 Sprachen auf der Welt. Wieviele es tatsächlich sind, weiß jedoch niemand, denn es gibt zum einen keine klare Scheidelinie zwischen „Sprache" und „Dialekt" (ist Schwyzerdütsch eine eigene Sprache, ein Dialekt oder eine Sammlung von Dialekten?) und zum andern weiß man schlichtweg zu wenig über die Sprachen der Welt. Das erste Problem ist grundsätzlicher Natur, es ist, als wollte man die Anzahl der Wolken an einem wolkigen Tag angeben. Das zweite hingegen ist eine Frage der empirischen Forschung und damit im Prinzip sehr wohl lösbar. Aber die Antwort verändert sich kontinuierlich, denn die meisten der Sprachen drohen in den nächsten vier bis sechs Generationen auszusterben – viele werden schon jetzt nur noch von älteren Menschen gesprochen. Durchschnittlich stirbt mindestens alle zwei Wochen eine Sprache, und um 2150 werden aller Voraussicht nach nur noch 700 bis 900 oder gar deutlich weniger Sprachen gesprochen werden. Mit jeder Sprache verschwindet unwiederbringlich viel kulturelles Wissen über die Welt und ein einzigartiger Gegenstand der Sprachwissenschaft.

Wieviele dieser – sagen wir jetzt einmal – 6.500 Sprachen sind denn überhaupt gut beschrieben? Das hängt natürlich davon ab, was man unter „gut beschrieben" versteht. Nimmt man einmal, ein sehr bescheidenes Kriterium, an, dass es dafür mindestens drei Werke zur Grammatik und drei zum Wortschatz der betreffenden Sprache geben muss, dann trifft dies auf schätzungsweise 100 Sprachen zu; es können aber auch 150 sein. Mit anderen Worten: Selbst nach diesem moderaten Maßstab sind allenfalls 2 Prozent aller Sprachen, die es derzeit gibt, gut beschrieben. Von den übrigen 98 Prozent wissen wir wenig, weniger, so gut wie nichts. Dieser unbefriedigende Stand hat vor allem zwei Gründe. Der eine ist die unerhörte Komplexität eines jeden sprachlichen Systems, so wie es eine Sprachgemeinschaft im Laufe vieler Jahre entwickelt hat. Es ist eine Illusion zu glauben, dass „primitive" Sprachen strukturell einfacher seien als etwa das Deutsche oder Englische – oft ist eher das Gegenteil der Fall. Lediglich der Wortschatz ist weniger umfangreich als bei den bedeutenden Kultursprachen. Der zweite Grund ist der Mangel an verlässlichen Daten. Bis vor 50 Jahren gab es außer Bleistift und Papier kaum Möglichkeiten, eine bislang unbekannte oder wenig erforschte Sprache – beispielsweise die der Eipo im Inneren Neuguineas – aufzuzeichen, um sie dann mit Sorgfalt zu studieren. Die einzige Möglichkeit, sich einigermaßen solide Kenntnisse von ihr zu verschaffen, bestand darin, dorthin zu fahren und einige Jahre dort zu leben. Daher stammen auch die meisten verfügbaren Wörterbücher und Grammatiken, die üblicherweise nach dem Vorbild des Lateinischen gestrickt sind, von frommen Missionaren, deren linguistische Expertise sich, anders als ihr Glaube, nicht selten in Grenzen hält.

So beruhen fast alle unsere Vorstellungen über die Natur der menschlichen Sprache auf einigen wenigen Exempeln wie Griechisch, Latein, Englisch, Deutsch, Chinesisch (dies schon weniger) und vielleicht zwei, drei



ELAN ist das wichtigste Werkzeug zur Aufbereitung und Analyse der Rohdaten

Dutzend weiteren. Das führt nicht nur zu einem einseitigen und verzerrten Bild von der Fülle möglicher sprachlicher Strukturen, sondern auch zu sehr unzulänglichen Vorstellungen davon, wie Sprachproduktion und Sprachverstehen – also die Sprache im tatsächlichen Gebrauch – funktionieren. Am deutlichsten schlägt sich die unbefriedigende Faktenlage vielleicht in dem für die Praxis so wichtigen Bereich des Sprachenlernens nieder. Der Erstspracherwerb des Kindes wie der Erwerb weiterer Sprachen im Kindes- oder Erwachsenenalter sind überaus komplexe Prozesse, die sich über viele Jahre erstrecken, eine erhebliche Variabilität zeigen und deren Verständnis, wenn es auf gut abgesicherten und wohlgegründeten Befunden beruhen soll, eine ungeheure Datenmenge und deren sorgfältige Analyse erfordern. Diese Daten haben wir aber allenfalls für einige wenige Sprachen, und auch dort sind sie selten so, dass man sie in großem Maßstab untersuchen kann; wir wissen einiges darüber, wie das Englische gelernt wird, wir wissen sehr wenig darüber, wie das Chinesische gelernt wird.

Dies ändert sich gerade rasant. Der erste Schritt war die Möglichkeit, Sprachdaten mit Audio- oder Videogeräten in hoher Qualität aufzuzeichnen. Von ganz besonderer Bedeutung ist diese Möglichkeit für die vielen Sprachen, die derzeit vom Aussterben bedroht sind – im Schnitt verschwindet alle zwei Wochen eine Sprache. Wenn diese Sprachen umfassend und verlässlich dokumentiert sind, möglichst nicht nur in Texten, sondern in Aufzeichnungen tatsächlicher

Kommunikation, dann leben sie als Teil des menschlichen kulturellen Erbes zumindest in dieser Form weiter: nicht genug, aber mehr als nichts. Der zweite und auf lange Sicht noch folgenreichere Schritt ist das Aufkommen digitaler Methoden, die all diese Daten nicht nur zu archivieren, sondern unter den verschiedensten Aspekten zu analysieren erlauben. Hier haben sich Möglichkeiten ergeben, von denen die Sprachwissenschaft vor einigen Jahren nur träumen konnte. So entstehen an verschiedenen Orten „digitale Sprachenarchive". Eines der größten, wenn nicht überhaupt das größte und technisch fortgeschrittenste, ist in den letzten 15 Jahren am Max-Planck-Institut für Psycholinguistik (Nijmegen) aufgebaut worden. Da es auch die gesamten Daten des von der Volkswagen-Stiftung geförderten Projektes „Dokumentation bedrohter Sprachen" (DoBeS) umfasst, hat es inzwischen eine Größe erreicht, die seine Fortführung im Rahmen eines Max-Planck-Instituts unmöglich macht. Deshalb ist es seit Oktober 2011 unter dem durchaus ambitionierten Namen „The Language Archive" (TLA) in die Obhut der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW), der Max-Planck-Gesellschaft (MPG) und der Königlich-Niederländischen Akademie der Wissenschaften (KNAW) übernommen worden und wird gemeinsam von den Partnern und dem MPI für Psycholinguistik, an das es organisatorisch angebunden ist, finanziert. Über die Weiterführung soll nach vier Jahren entschieden werden. Geleitet wird es von Wolfgang Klein und einem Steering Board, dem je ein Vertreter der BBAW, der KNAW und der MPG angehören.

Fotos: Gunter Senft

Das Kilivila auf den Trobriand-Inseln, berühmt geworden durch die Forschungen Malinowskis, wird derzeit noch von etwa 20.000 Menschen gesprochen

Die Bestände des TLA umfassen derzeit etwa 80 Terabyte an durch Metadaten aufbereiteten und damit durchsuchbaren Daten aus circa 200 Sprachen; dies entspricht etwa 20.000 Stunden an Video- und Audiomaterial sowie einer schwer quantifizierbaren Menge reiner Textdaten; derzeit kommen jährlich etwa 12 Terabyte hinzu. Die Daten entstammen im wesentlichen drei Quellen. Dies sind zum

Höchstens 2 Prozent aller Sprachen, die es derzeit gibt, sind gut beschrieben.

ersten die über Jahre hinweg aufgebauten Bestände am MPI für Psycholinguistik. Die zweite Gruppe bilden die Materialien des schon genannten DoBeS-Projekts der Volkswagen-Stiftung, das 2000 begonnen wurde und voraussichtlich bis 2016 läuft; bislang haben 60 kleine Teams von Forscherinnen und Forschern etwa 80 vom

Aussterben bedrohte oder inzwischen gar schon ausgestorbene Sprachen dokumentiert. Die dritte Gruppe ist heterogen, es sind dies Materialien, die dem TLA meist als – oft noch nicht einmal digitalisierte – Rohdaten von verschiedenen Forscherinnen und Forschern oder Institutionen zum Sichern und Aufbereiten übergeben wurden. Diese Gruppe soll in den kommenden Jahren systematisch ausgeweitet werden.

Eines der größten, wenn nicht überhaupt das größte und technisch fortgeschrittenste Sprachenarchiv

Für die Forschung von Nutzen sind all diese Daten nur, wenn Werkzeuge bereitgestellt werden, die es erlauben, die Materialien zu archivieren, mit Annotationen zu versehen, nach verschiedenen Gesichtspunkten zu durchsuchen und – nach Möglichkeit automatisch – zu analysieren. Dazu hat das TLA ein unter dem Namen LAT („Language Archiving Technology“) zusammengefasstes Bündel an Softwaretools entwickelt, die frei verwendbar sind und von der Forschung weltweit auch bereits viel genutzt werden. Dafür seien hier zwei Beispiele genannt: Die Sprachdaten liegen in der Regel zunächst als Audio- beziehungsweise Videomaterial vor, das als erstes in einzelne „Sessions“ unterteilt und mit entsprechenden Metadaten versehen wird; sonst kann man gar nichts darin finden. Auf diese Weise sind die 20.000 Stunden in insgesamt 73.000 „Sessions“ aufgeteilt, die aber in sich auch zunächst nur einmal Audio- und Videomaterial sind. Dies muss nun linguistisch annotiert, also transkribiert und mit allen möglichen Angaben zur Morphologie, zur Syntax, zur Intonation, zum Wortschatz, zur Gestik, Mimik und dergleichen mehr versehen werden. Erst dann kann man wirklich eine Beschreibung der Sprache

Dies sind die bisher im DoBeS-Projekt dokumentierten Sprachen

oder des Kommunikationsverhaltens in Angriff nehmen. Diese linguistische Annotation ist extrem aufwendig, aber unabdinglich, und je reicher sie ist, umso mehr kann man damit tun. Sie kann in der Regel nicht mehr von den Mitarbeiterinnen und Mitarbeitern des TLA selbst geleistet werden, sondern man braucht dazu Experten für die betreffende Sprache; zumeist sind dies jene, die die Daten aufgenommen haben. Das TLA hat dazu ein sehr flexibles, leicht handhabbares und effizientes Werkzeug namens ELAN entwickelt, das allmählich zum Standard für solche linguistische Annotationen wird. ELAN stützt sich auf die Expertise der Linguistinnen und Linguisten, die sich die Daten anschauen und ihre Entscheidungen treffen. Das ist das einzig verlässliche Vorgehen, es ist aber überaus zeitraubend. Daher wäre es eine eminente Erleichtung, wenn man die linguistisch relevanten Muster in den Audio- und Videodaten automatisch oder zumindest teilautomatisch erkennen würde.

Das TLA entwickelt dazu gemeinsam mit zwei Fraunhofer-Instituten (IAIS, St. Augustin, und HHI, Berlin) ein als AVATecH („Advanced Video and Audio Technology in Humanities Research") bezeichnetes Programmpaket, das dies in gewissen Grenzen leistet. Eine perfekte Lösung ist derzeit außer Reichweite und wird auch nicht angestrebt, aber die Arbeit lässt sich damit massiv beschleunigen.

Das TLA wird aus öffentlichen Mitteln finanziert und soll daher auch der gesamten Öffentlichkeit zugänglich sein; dies gilt sowohl für die Daten wie für die Werkzeuge. Für erstere gibt es jedoch juristische und ethische Beschränkungen, die sich zum einen aus den Rechten derer ergeben, die die Daten aufgenommen haben, zum anderen aus den Persönlichkeitsrechten derer, die aufgenommen wurden. Dies kann im Einzelfall sehr unterschiedlich aussehen; so übertragen manche Urheber ihre Rechte auf das Archiv, andere behalten sich die Entscheidung über die Nutzung vor. Das TLA hat dazu ein gestuftes Zugangssystem entwickelt, das von „völlig frei" bis zu „derzeit nur den Urhebern / den aufgenommenen Personen zugänglich" reicht. Angestrebt wird immer eine maximale Offenheit.

Das TLA ist in ein Netzwerk von Kooperationen eingebunden. Daran sind zunächst einmal andere Einrichtungen oder Projekte der drei Partner beteiligt, an der BBAW etwa das Zentrum Sprache und hier insbesondere das „Digitale Wörterbuch der deutschen Sprache" und das „Deutsche Textarchiv" sowie TELOTA (The electronic life of the Academy), an der KNAW das Meertens Instituut. Bereits genannt worden ist das DoBeS-Projekt der Volkswagen-Stiftung sowie das AVATeCH-Projekt mit der Fraunhofer-Gesellschaft. Von ganz besonderer Bedeutung sind jedoch auch übergreifende Initiativen, insbesondere die von der EU getragenen Netzwerke CLARIN („Common Language Resources and Technology Infrastructure") und DASISH („Data Service Infrastructure for the Social Sciences and Humanities"). Nicht zuletzt lebt das TLA jedoch von der Kooperation mit vielen einzelnen Sprachforscherinnern und -forschern aus aller Welt, die oft auf sich gestellt ihren Beitrag zu einer besseren Erforschung der menschlichen Sprachvielfalt und damit zu einem besseren Verständnis dessen beitragen, was die menschliche Sprache ausmacht.

Prof. Dr. Wolfgang Klein *ist Direktor am Max-Planck-Institut für Psycholinguistik in Nijmegen und Projektleiter von „The Language Archive". Er ist Mitglied der Berlin-Brandenburgischen Akademie der Wissenschaften und leitet dort das Zentrum Sprache.*

→ *www.mpi.nl/tla*
→ *www.mpi.nl/dobes*