Zur gegenseitigen Befruchtung

In einem neuen Forschungsprojekt der Universität Würzburg arbeiten Sprachwissenschaftler, Informatiker und Bioinformatiker zusammen. Ihre Idee: Wenn sich Sprachen ähnlich wie Arten entwickeln, müssten sich die Arbeitsmethoden aus allen drei Bereichen gemeinsam nutzen lassen.

Die Idee klingt verlockend: So, wie sämtliche Lebewesen einen gemeinsamen Vorfahren haben, lassen sich auch Sprachen auf einen Ursprung zurückführen. Dann sollten die Gesetze der Evolution nicht nur für Bakterien, Blumen, Bienen, sondern auch für Wörter, Sätze und Sprachen gelten.

Parallelen lassen sich in hinreichender Zahl finden: Während die Sprachwissenschaftler zwischen zehn und 80 grundlegende Lautsystemeinheiten der Sprache kennen, so genannte Phoneme, zählen die Biologen 20 Aminosäuren als elementare Bausteine des Lebens. Den geschätzten 5000 bis 10.000 übergeordneten Einheiten in der Sprache, den Bedeutung tragenden Einheiten, entsprechen in der Biologie die so genannten Domänen, längere Ketten von Aminosäuren, die bestimmte Funktionen erfüllen. Von ihnen soll es ebenfalls um die 5000 geben. Und den vermuteten 100.000 Proteinen stehen über den Daumen gepeilt Wörter in gleicher Anzahl gegenüber – lässt man zusammengesetzte Konstrukte wie die „Donaudampfschifffahrtskapitänswitwe“ mal außen vor.

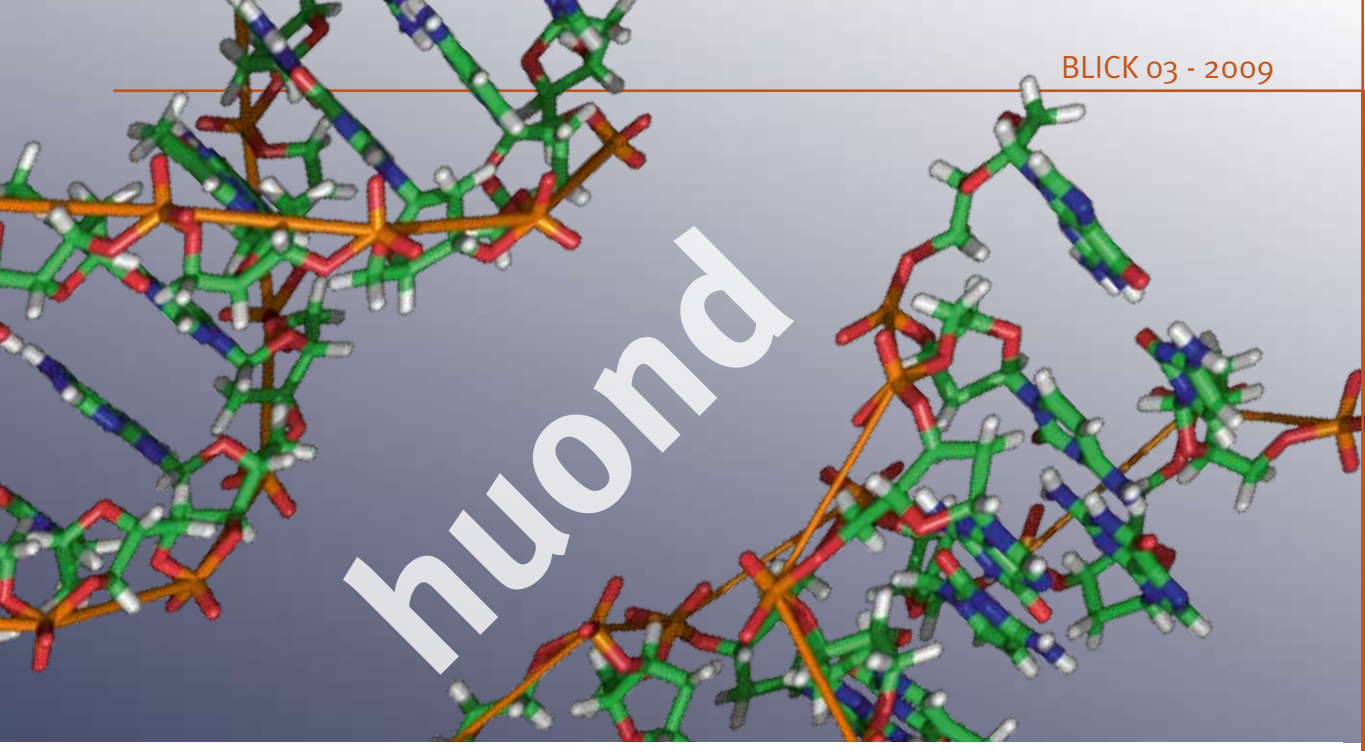
Selbst auf anderen Ebenen lassen sich Analogien zwischen Sprach- und Artentwicklung erkennen: Die Verwandtschaft des Italienischen mit dem Französischen entspräche in etwa der von Löwe und Tiger. Unterschiedliche Hunderassen: Sie wären vergleichbar mit Dialekten wie Schwäbisch, Bairisch oder Sächsisch. Tatsächlich kann man sowohl für Sprachen als auch Lebewesen Stammbäume zeichnen, die sich verblüffend ähnlich sehen.

Schon Darwin vermutete Parallelen zwischen Arten und Sprachen

Wenn diese Ähnlichkeit zwischen zwei so unterschiedlichen Fachgebieten tatsächlich existiert: Sollten sich dann nicht die Verfahren, mit denen Biologen ihre gewaltigen Datenmengen verarbeiten und verwalten, und die Methoden, mit denen Sprachwissenschaftler Verwandtschaftsbeziehungen und historische Entwicklungen aufspüren, wechselseitig nutzen lassen? Zum gegenseitigen Vorteil aller Beteiligter? Etwas in der Art probiert ein neues Forschungsprojekt an der Universität Würzburg. Daran beteiligt sind Sprachwissenschaftler, Informatiker und Bioinformatiker.

„Die Idee ist ja nicht neu: Schon Darwin hat vor 150 Jahren vermutet, dass sich Sprachen ähnlich entwickeln wie Arten. Wir wollen das jetzt genauer untersuchen“, sagt Jörg Schultz, Professor am Lehrstuhl für Bioinformatik der Universität Würzburg. Dabei gehe es allerdings nicht um Organismen und Populationen. Der Biowissenschaftler will auf molekularer Ebene tätig werden und Gene, Gensequenzen und Genome in das Projekt mit einbeziehen. Von der Zusammenarbeit mit den Geisteswissenschaftlern verspricht sich Schultz Erkenntnisse, die er auch in seiner Arbeit nutzen kann. „Gensequenzen, Aminosäuren, Proteine: Wir haben in den vergangenen Jahren einen gewaltigen Berg an Daten gesammelt“, sagt Schultz.

Das Problem dabei bringt Dietmar Seipel, Professor am Lehrstuhl für Informatik I der Universität Würzburg, auf den Punkt: „We are drowning in



data, but starving for knowledge“. Bei der Suche nach diesem Wissen können sich Bioinformatiker, Informatiker und Sprachwissenschaftler gegenseitig befruchten. Die Informatiker und die Bioinformatiker verfügen über Methoden zur Auswertung und Interpretation von solchen Massendaten. „Die Informatik arbeitet seit Langem an Methoden zum Auffinden von Mustern und Regelmäßigkeiten in großen Datenmengen, dem sogenannten Data Mining. Außerdem wurden bereits Methoden untersucht, um Daten aus unterschiedlichen Quellen beziehungsweise Bereichen aufeinander abzubilden, das sogenannte Ontology Alignment“, so Seipel. Und wenn denn tatsächlich die Parallelen zwischen der Entstehung der Arten und der Sprachen so groß wie erhofft sind, dann könnten das Wissen der Sprachforscher und die Methoden der Bioinformatiker und Informatiker vielleicht sogar einen Blick in die Vergangenheit der Arten ermöglichen.

Hier kommt Werner Wegstein ins Spiel. Wegstein war von 1975 an am Aufbau der neuen sprachwissenschaftlichen Abteilung am Institut für deutsche Philologie beteiligt. Sein Interesse am Einsatz von Computern und Methoden der Informatik in der Sprachwissenschaft zeigte sich schon in seiner Habilitationsschrift; die drehte sich um „Texte im Datennetz. Bausteine zu einer computergestützten Philologie“. Von 2003 bis 2008 hatte Wegstein die Professur für EDV-Philologie inne. Aber wie kann der Sprachwissenschaftler Einsichten des Biologen stützen?

„Es gibt gute Gründe dafür, anzunehmen, dass der gemeinsame Vorfahre der indogermanischen Sprachen vor etwa 6000 Jahren relativ einheitlich war. Daraus hat sich in der Zeit danach die Familie indogermanischer Sprachen entwickelt, die man heute auf einem Großteil der Erde spricht“, sagt Wegstein. Anders ausgedrückt: Innerhalb von gerade einmal 200 bis 300 Generationen hat sich aus einer Ursprungssprache die heutige Vielfalt indogermanischer Sprachen entwickelt, mit zahlreichen Unterfamilien wie beispielsweise den keltischen und germanischen Sprachen im Westen und Norden, den italischen und romanischen Sprachen im Süden, den baltischen und slawischen Sprachen im Osten und weit im Fernen Osten die große Familie der indo-iranischen Sprachen. Der Sprachwandel beweist also eine hohe Dynamik. Für die germanischen Sprachen, genauer für das Deutsche, lässt er sich zudem über die vergangenen rund 1000 Jahre einigermaßen zurückverfolgen, da es aus dieser Zeit hinreichend schriftliche Zeugnisse gibt.

Hilfe beim Blick in die Vergangenheit

Und was bedeutet das für die Biologie? „Wenn die Sprachwissenschaft innerhalb eines Zeitraums von 6000 Jahren die Entwicklung der letzten 1000 überblicken kann, und wenn man davon ausgeht, dass Sprachen und Arten sich auf ähnliche Weise entwickeln, heißt das in unseren Dimensionen beispielsweise: Der gemeinsame Vorfahre von Mensch und Maus hat vor rund 60 Millionen Jahren gelebt. Mit dem Wissen der Sprachwissenschaftler könnten wir also Schlüsse über die vergangenen zehn Millionen Jahre ziehen“, erklärt Jörg Schultz.

Oder konkreter: Die Bioinformatik ist nicht in der Lage, einen Blick in die Vergangenheit zu werfen – gut erhaltenes Erbmaterial aus der Zeit der Dinosaurier gibt es nur in *Jurassic Park*, nicht aber in der Realität. Bioinformatiker können die Historie nur extrapolieren und sind dann nicht in der Lage zu überprüfen, ob die Vergangenheit tatsächlich so ausgesehen hat, wie sie sich das vorstellen. Die Hilfe der Sprachwissenschaft funktioniert so: „Linguisten könnten die Methoden, mit denen wir arbeiten, auf heutige Texte anwenden und das Ergebnis mit realen Beispielen aus der Vergangenheit vergleichen“, erklärt Schultz. Stimmt das Ergebnis wäre dies zumindest ein wichtiges Indiz dafür, dass die bioinformatischen Methoden so ganz falsch nicht sein können.

Umgekehrt erwachsen natürlich auch den Sprachwissenschaftlern Vorteile aus der Zusammenarbeit: Sie bringen ihre Daten jetzt in eine elektronisch auswertbare Form und können dann etablierte bioinformatische Methoden auf diese Daten anwenden. Das ist die Schnittstelle des Projekts, an der Esther

Ratsch, Diplom-Biologin am Institut für Bioinformatik, zur Zeit arbeitet: eine Ontologie, in der die Analysestrukturen der Sprachwissenschaft mit denen der Bioinformatik in Beziehung gesetzt werden. „Wenn es tatsächlich vergleichbare Mechanismen sind, die die Evolution und die Sprachentwicklung treiben, sollten wir auf diesem Weg in den sprachwissenschaftlichen Daten Muster finden, die die Linguisten selbst noch nicht entdeckt haben", hofft Jörg Schultz.

Geistes- und Naturwissenschaft miteinander kombinieren

Und welche Rolle spielen die Informatiker in diesem Forschungsprojekt? „Die Informatik arbeitet an Methoden zur Sprachverarbeitung, sowie an Techniken zur Verwaltung von großen Datenmengen und zur Erkennung von Mustern und Zusammenhängen in diesen Datenmengen", erklärt Dietmar Seipel. Diese Methoden und Techniken sollen im Laufe des Projektes auf die sprachwissenschaftlichen Daten angewendet und bei Bedarf erweitert und verallgemeinert werden, um der Datenflut Herr zu werden. Nach der Aufbereitung der sprachwissenschaftlichen Daten wird die Informatik auch versuchen einen Beitrag zur Gegenüberstellung der sprachlichen und biologischen Begriffe, dem „Alignment der Ontologien", zu leisten.

Auslöser der ungewöhnlichen Zusammenarbeit in dem Projekt mit dem umfänglichen Titel „Wechselwirkungen zwischen linguistischen und bioinformatischen Verfahren, Methoden und Algorithmen: Modellierung und Abbildung von Varianz in Sprache und Genomen" war ein Aufruf des Bundesforschungsministeriums: Dort waren förderungswürdige Projekte gesucht, die Geistes- und Naturwissenschaften miteinander kombinieren und in Wechselwirkung zueinander setzen. Ein Aufruf ganz nach dem Geschmack von Werner Wegstein. Der stand mit den Informatikern eh schon in regem Kontakt; und sowohl für den Informatiker wie für den Sprachwissenschaftler war die Bioinformatik der ideale Forschungspartner aus dem Bereich der Naturwissenschaften. Die Idee von ähnlichen Entwicklungen bei Sprache und Arten hatten die Beteiligten schon öfter diskutiert.

Als weitere Forschungspartner brachten die Wissenschaftler am Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften der Universität Trier eine Fülle historischer Wörterbuchdaten in das Projekt mit ein, und das Institut für Deutsche Sprache in Mannheim steuerte die gegenwartssprachlichen Wortschatzdaten bei. Mit Erfolg: „Von mehr als 60 Gruppen, die sich beworben hatten, kamen 13 ins Auswahlverfahren und wurden zur Präsentation nach Bonn eingeladen. Unser Projekt wurde als eines von acht Projekten als förderungswürdig eingestuft", sagt Wegstein.

> „We are drowning in data, but starving for knowledge"
>
> Dietmar Seipel

Was jetzt geschieht, klingt nach viel Fleißarbeit. „Damit wir Wörter mit biologischen Strukturen vergleichen können, brauchen wir zunächst eine adäquate Datenbank sprachlicher Strukturen", erklärt Esther Ratsch. Oder, in der Fachsprache: eine Metalemmaliste. Eine solche Liste dient den Sprachwissenschaftlern quasi als Steinbruch; dort sollen sie einmal sämtliche Grundbausteine der deutschen Sprache finden können. Als Ausgangsmaterial dienen zunächst die 30.000 häufigsten Wörter des Standarddeutschen aus einer Datensammlung von rund 3,5 Milliarden Wörtern. Diese sollen dann in einer Datenbank mit historischen und mit Dialektwörterbüchern verknüpft werden. „Damit erhalten wir eine Art dreidimensionalen Zugang zum Wortschatz, sowohl in der Zeit als auch in der Fläche", erläutert Wegstein.

Wie diese Verknüpfung geht? Wal-Mart hat's vorgemacht. „Die Supermarktkette hatte als Erster die Idee, die Warenkörbe ihrer Kunden zu analysieren", erklärt Christian Schneiker, Informatiker und Doktorand in dem Projekt. Aus der Erkenntnis, dass Kunden, die Windeln kaufen, häufig auch Bier und Chips mitnehmen, entstand die Idee, diese Produkte möglichst nah beieinander zu platzieren. „Das waren die Anfänge des so genannten Data Minings, einer Technik, mit der wir jetzt die Sprachentwicklung zu erfassen versuchen", ergänzt Dietmar Seipel. Mit dem Unterschied, dass es dabei nicht um die Beziehung von Knabbersachen und Babybedarf geht, sondern um die Beziehungen zwischen dem mittelhochdeutschen „Hunt", dem lothringischen „Hond" und dem neuhochdeutschen „Hund". „Vereinfacht ausgedrückt lassen wir einen Algorithmus auf den Datenbestand los und schauen dann mal nach, ob er sinnvolle Muster und Regelmäßigkeiten in den Daten findet", sagt Christian Schneiker.

Permanente gegenseitige Erklärungen sind nötig

Dass das gar nicht so einfach ist, wenn Fachleute aus drei unterschiedlichen Fachgebieten miteinander auskommen wollen, haben die Projektbeteiligten schnell gemerkt. „Biologen und Sprachwissenschaftler müssen sich permanent gegenseitig erklären, wie sie denken und was sie meinen. Und die Informatiker müssen beide Seiten verstehen", sagt Jörg Schultz. „Außerdem sind wir gezwungen, sehr sauber zu definieren, worüber wir gerade sprechen, damit das beim Gegenüber auch richtig ankommt", ergänzt Werner Wegstein.

Als „spannenden Bereich der Grundlagenforschung" empfinden alle Beteiligten dies Projekt. Der interdisziplinäre Ansatz habe schon jetzt Anstoß für jede Menge neuer Ideen und Anregungen für die eigene Arbeit gegeben, sagen sie. Und was, wenn am Ende herauskommen sollte, dass sich die Entwicklung der Sprachen doch nicht mit der der Arten vergleichen lässt? „Das ist dann Wissenschaft", sagt Jörg Schultz. Wenn sich eine Theorie als falsch herausstellt, gelte es, eine neue zu entwickeln. Und auch dann sei die Arbeit nicht umsonst gewesen, stimmen Dietmar Seipel und Werner Wegstein überein: „Die Metalemmaliste haben wir auf alle Fälle."

Gunnar Bartsch