Stringbasierte Algorithmen zur Rekonstruktion von Sprachverwandtschaften

Exposé zur Studienarbeit

Mirko Hochmuth

22. Juli 2004

Betreuer: Prof. Ulf Leser
(HU Berlin, Inst. für Informatik)
Prof. Anke Lüdelig
(HU Berlin, Inst. für Deutsche Sprache und Linguistik)

Zeitrahmen: 09. Juli - 08. Oktober 2004

Motivation

Im Jahr 2003 wurde an der Humboldt Universität zu Berlin der "Interdisziplinäre Forschungsverbund Linguistik - Bioinformatik" gegründet. Sein Ziel ist es, Methoden und Erkenntnisse aus dem Bereich Biologie/Bioinformatik auf den Bereich Linguistik/Korpuslinguistik zu übertragen und umgekehrt. Beide Forschungsgebiete stehen vor der Aufgabe, große Mengen von Zeichenketten zu analysieren und zu verarbeiten [1].

In der Bioinformatik wird unter anderem versucht, mittels *phylogenetischer Algorithmen* Verwandtschaftsbeziehungen von Spezies zu finden. Diese Algorithmen werten große Mengen von Strings (DNA- bzw. Gensequenzen) aus und bestimmen den genetischen Abstand von Spezies. Über diesen Abstand lassen sich dann Rückschlüsse auf Herkunft und Abstammung der untersuchten Spezies ziehen und Artenstammbäume konstruieren.

Eine Teilaufgabe der Linguistik ist es, die *Verwandtschaftsbeziehungen von Sprachen* (z.B. die Verwandtschaften innerhalb der indoeuropäischen Sprachenfamilie) zu untersuchen.

Bisher wurden diese Sprachverwandtschaften größtenteils in mühevoller Handarbeit rekonstruiert, die Idee liegt also nahe, die in der Bioinformatik bewährten Algorithmen zur Bestimmung von Artverwandtschaften auf Sprachen anzuwenden.

Zielsetzung

Ziel dieser Studienarbeit ist es, im Rahmen des Forschungsverbundes, Möglichkeiten zu finden, stringbasierte phylogenetische Algorithmen aus der Bioinformatik auf die Bestimmung von Sprachverwandtschaften zu übertragen. Eine der gefunden Möglichkeiten soll an konkreten linguistischen Problemen/Daten erprobt werden.

Herangehensweise

Zunächst erfolgt ein eingehendes Literaturstudium zur Einarbeitung in die linguistischen Fragestellungen (Sprachverwandtschaften, Phonetik, Lautschrift, Lautverschiebungen, ...) sowie in phylogenetische Algorithmen. Bisherige Arbeiten zur Kombination der beiden Themengebiete werden genauer untersucht [u.a. 2, 3].

Phylogenetische Algorithmen setzten ein *Abstandsmaß* für die zugrundeliegenden Einheiten (Arten, Sprache) voraus. Es ist also eine Charakterisierung der verschiedenen Ebenen, auf denen man dieses Abstandsmaß berechnen kann (Wörter, Wörter in Lautschrift, Sätze, Texte, ...) zu erstellen und eine Diskussion zu führen, welche Ebene sich am besten für die Aufgabe eignet und welche sich im gesetzten Zeitrahmen umsetzten läßt. Anschließend soll ein Verfahren ausgewählt und implementiert werden.

Als Datenbasis für die gewählte Methode dienen verschiedene Versionen des *Vater unser* sowie eine Geschichte der Bibel in verschiedenen Sprachstufen des Deutschen (u. a. Altsächsisch, Althochdeutsch, Mittelhochdeutsch, Frühneuhochdeutsch und Neuhochdeutsch).

Zur Berechnung eines Stammbaums müssen Lösungen für folgende Teilprobleme gefunden werden:

- Zuordnung der Wörter (Wortalignment)
- Definition eines Abstandsmasses für Wörter
- Definition einer Substitutionsmatrix für Zeichen (z. B. ist es weniger teuer, ein i durch ein e zu ersetzen als durch ein o oder gar einen Konsonanten)
- Verknüpfung der Ähnlichkeiten verschiedener Wortpaare zu Ähnlichkeiten der Texte
- Anwendung eines phylogenetischen Verfahrens zur Berechnung des Stammbaums

Quellen und Literatur

1. http://korpling.german.hu-berlin.de/forschungsverbund/

2. Russell D. Gray & Quentin D. Atkinson: Language-tree divergence times support the Anatolian theory of Indo-European origin. In: Nature Vol. 426, S. 435-439, 2003

3. Wilbert Heeringa: Measuring Dialect Pronunciation Differences using Levenshtein Distance. Thesis at Rijksuniversiteit Groningen, 2004