Der Einsatz von PCs mit UNIX-like Tools zur Sprachforschung

SHIROOKA Keiji (jjksiro@hss.shizuoka.ac.jp)

Wollte man früher eine computerunterstützte Forschungsarbeit durchführen, arbeitete man mit einem größeren Computer im Computerzentrum. Man mußte dabei entweder selbst mit einer schwierigen Programmiersprache programmieren können, oder man beauftragte Fachleute mit dieser Aufgabe. Infolge der Entwicklung von PCs (Personal Computer) gab es dann einige lexikologische und graphematische Pilotstudien mit PC und BASIC. BASIC ist zwar relativ leicht zu erlernen, aber doch nicht für Laien geeignet. Nun sind die PCs soweit fortgeschritten, daß auch Programmiersprachen wie BASIC zumindest für einige Bereiche unnötig geworden sind. Man kann jetzt mit sogenannten UNIX-like Tools ziemlich viele Aufgaben lösen.

1. UNIX-like Tools

Als UNIX-like Tools bezeichne ich alle Werkzeuge, die ursprünglich aus UNIX stammen und auf verschiedene Weise Textdateien verarbeiten (sie können auch Text Tools, DOS Tools, UNIX-style Utilities heißen). Die Werkzeuge sind einzeln oder kombiniert einzusetzen.

1.1 Sortieren mit SORT

Mit SORT kann man Zeilen einer Textdatei auf verschiedene Weise sortieren. Zu sortieren sind sowohl Zahlen als auch normale Schriftzeichen. Sortierte Zahlendaten können z. B. Grundlage der weiteren Analyse sein. Ich habe als Beispiel dafür Zahlendaten von „furniture“ und „Möbel“ sortiert und die daraus gewonnenen Fakten analysiert.

1.2 Frequenzlisten

Mit SORT, UNIQ, CUT, FOLD, REV, WORD, AWK kann die Häufigkeit der Wörter, Graphketten und Schriftzeichen untersucht werden.

1.3 Listenvergleich

Listen kann man mit COMM mühelos vergleichen. Ich habe mit COMM zwei Listen von Wörtern, die jeweils als Grundwortschatz für Deutsch als Fremdsprache vorgeschlagen wurden, verglichen.

2. Textrecherche mit GREP

GREP und Reguläre Ausdrücke dienen im allgemeinen dazu, in einer Textdatei Graphketten in der Größenordnung von Schriftzeichen bis Phrasen zu suchen. Eine Art GREP ist KKC, und es ermöglicht sogar, die Fundstellen in KWIC-Format (Keyword in Context) übersichtlich auszudrücken.

3. Untersuchungsbeispiele mit AWK

Mit AWK lassen sich auch schwierigere Aufgaben lösen. Zur Bearbeitung einer Wortliste von mehr als 55000 Wörtern, die ursprünglich aus drei verschiedenen Wörterbüchern stammen, habe ich AWK eingesetzt. Zwei Beispiele für die Anwendung waren:

3.1 Sprachdidaktisch wichtige Graphketten

Im Deutschunterricht lernt man in der Einführung, wie man Graphketten wie „ch", „sch", „st", „ck", „eu", „dt" ausspricht. Müssen alle Graphketten, die nicht dem Schriftbild entsprechend ausgesprochen werden, schon im Anfängerunterricht durchgenommen werden? Gibt es keine Graphketten, die häufiger und wichtiger sind? Mit AWK habe ich die Häufigkeit der insgesamt 41 Graphketten, die aus mehr als zwei Schriftzeichen bestehen, untersucht. Als Material habe ich nicht nur die Stichwortliste der drei verschiedenen Wörterbücher, sondern auch eine Wortliste aus dem SPIEGEL mit ca. 93 000 Wörtern benutzt.

3.2 Graphematische Minimalpaare: [l/r] und [b/w]

Mit AWK kann man alle graphematischen Minimalpaare aus einer Wortliste, in der jedes Wort zeilenweise steht, herausnehmen. Wenn man z. B. Minimalpaare mit der Unterscheidung von „l" und „r" an der zweiten Stelle sucht, kann man folgenderweise vorgehen: Mit dem Regulären Ausdruck /^.[lr]/ kann man Zeilen mit „l" oder „r" als zweitem Schriftzeichen selektiv verarbeiten. Man eliminiert

dann das zweite Schriftzeichen, registriert den Rest des jeweiligen Wortes mit „rest= substr($0, 1, 1) substr($0, 3)" und zählt mit „arr[rest]++". Jede Form ohne das zweite Schriftzeichen wird gezählt. Wenn eine Form am Ende zweimal vorkommt, bedeutet es, daß es da Minimalpaare wie „Blockhaus" und „Brockhaus", „blühen" und „brühen", „flau" und „Frau" gibt.

Meiner Erfahrung nach ist es nicht leicht, ohne Vorkenntnisse Gebrauchsanweisungen zu verstehen. Man muß vorher wissen, was man überhaupt mit einem Gerät oder einem Werkzeug machen kann. Mein Ziel ist es daher, einen groben Überblick über das, was man mit einem PC und UNIX-like Tools machen kann, zu geben. Ich habe deshalb versucht, die Leistungsfähigkeit der Werkzeuge mit meinen konkreten Beispielen zu beweisen und damit auch mögliche Einsatzbereiche von PCs zur Sprachforschung zu zeigen. Aber ich habe nicht versucht, Gebrauchsanweisungen für die einzelnen Werkzeuge zu schreiben. Dies ist in einem Aufsatz von diesem Umfang auch nicht möglich.