Erstellung großer Aussprachelexika

Andreas Mengel

Zusammenfassung

Bei der Erstellung von Aussprachelexika sind neben der Festlegung der abzubildenden Varietät, des Transkriptionsalphabets und der Auswahl des transkribierten Vokabulars eine Reihe weiterer Aspekte zu berücksichtigen. Im EG-Forschungsprojekt ONOMASTICA mit der Aufgabe, für die im jeweiligen Land vorkommenden Eigennamen (Vor-, Nach-, Straßen-, Orts- und Firmennamen) Transkriptionen zu erstellen, wurden diesbezüglich Erfahrungen gesammelt, deren Essenz empfehlend dargestellt wird.

1 Bedarf an Transkriptionen

Für alle Zwecke, in denen prinzipiell unendlich große Vokabulare lautsprachlich zur Verfügung stehen müssen, sind abstrakte, symbolische Repräsentationen dieser Daten unverzichtbar. Dies gilt solange, wie die Aufnahme der Sprachsignale dieser Vokabulare die jeweils verfügbaren technischen Möglichkeiten übersteigen. Werden also Repräsentationen unterhalb der Wortebene benötigt, die mehr oder weniger leicht und erfolgreich erzeugt bzw. reproduziert und dann zusammengesetzt werden können, wird eine deren Struktur beschreibende Symbolsprache wie das IPA erforderlich.

Die für weite Teile gängige Zugriffssprache zur Repräsentation von Wörtern in unserer Alltagswelt ist unsere Orthographie, mithin ein Mittel, deren Struktur vor allen Dingen ein Kriterium vermissen läßt: Die kontextunabhängige 1:1-Entsprechung der Buchstaben zu Lauten. Aus diesem Grund ist es nötig, für jedes Wort des gewünschten Vokabulars in der orthographischen Variante eine Repräsentation in Lautschrift zu generieren.

2 Online-Transkription vs. Lexikon

Wenn Transkriptionen für orthographische Wörter zur Verfügung stehen sollen, so besteht entweder die Möglichkeit, ein System zu entwerfen, welches die orthographische in phonetische Information umsetzt, oder eine Liste, in der die jeweiligen Entsprechungen einzeln aufgeführt sind, zu erstellen. Verschiedene Kriterien sind für die Entscheidung zugunsten einer der beiden Alternativen zu nennen.

2.1 Benutzungsaufwand

Unter Benutzungsaufwand sind Resourcen zu verstehen, die während der Anfrage an ein System zur Verfügung stehen müssen. Als Kriterien sind hier Rechenaufwand und Speicherbedarf einerseits und Wartezeit des Benutzers andererseits zu nennen. Regeln brauchen weniger Speicher als ein Lexikon, der Rechenaufwand für die Berechnung der Transkription ist höher als das Suchen auf einer Platte; das Verhältnis der Wartezeiten ist wegen der mechanischen Trägheit der Speichermedien umgekehrt dazu. Diese Parameter jedoch sind stark vom jeweiligen Stand der Technik abhängig und damit in zunehmendem Maße vernachlässigbar.

2.2 Erstellungsaufwand

Erstellungsaufwand sind die Resourcen, die für die Erstellung des Lexikons oder des Systems gebraucht werden.

Für die Erstellung jedes Systems unabdingbar ist eine Datenbasis, mit Hilfe derer das System - auf welche Weise auch immer - trainiert wird. Weiter kann Hintergrundwissen mehr oder weniger stark vonnöten sein (s.a. 2.5). Insbesondere aber bei der Erstellung eines Regelwerkes ist der Aufwand proportional zu der Unregelmäßigkeit des Verhältnisses der orthographischen Struktur zur phonetischen und der Ingeniösität des Systementwicklers bei der Erfindung von Prozeduren, die diese Unregelmäßigkeiten dennoch abbildbar machen.

Es wird generell wird davon ausgegangen, daß die Formulierung von Regeln, deren Anzahl sich um Potenzen von der der Einzeleinträge, die transkribiert werden sollen, unterscheidet, eine schneller zu bewältigende Aufgabe ist. Dies ist der Fall, wenn die Struktur der jeweiligen Orthographie eindeutig auf alle Informationen der Lautung schließen läßt. Weil aber die Orthographie diesbezüglich zu idiosynkratisch ist, und deshalb eine eine Transkription verlangt wird, ist anzunehmen, daß für eine große Anzahl von Wörtern der Einsatz eines automatischen Transkriptionssystems ein annehmbares Resultat ergeben wird, eine andere Gruppe von Fällen jedoch nur unzureichend umzusetzen sein wird. Weiter gibt es während des Trainings oder der Implementierung eines Systems nur eine Methode, die von diesesm System falsch transkribierten Wörter zu entdecken: Den Vergleich des Outputs mit den Transkriptionen in einer schon bestehenden verläßlichen Liste. Diese Liste muß aber vorher erstellt worden sein und wird, wenn das System an ihr gemessen wird, immer mehr Einträge beinhalten (müssen) als die Regeln richtig transkribieren können, womit sich, wenn eine sehr niedrige Fehlerrate angestrebt wird, ein automatisches Transkriptionssystem für die oben beschriebene Aufgabe als unbrauchbar erweist. Um dieser Problematik auszuweichen, wird vielfach vorgeschlagen, einem Transkriptionssystem eine Abfrage einer Ausnahmeliste vorzuschalten. Diese Liste widerum erfüllt ihren Sinn nur für bekannte schwierige Wörter. Problematische Fälle an sich sind so nicht zu erkennen.

Phonetisches Wissen wird für die Erstellung eines Lexikons nötig, wenn die Struktur der Transkriptionen insgesamt einheitlich sein soll. Ansonsten ist nur konkretes Wissen um die jeweilige Aussprache der Einträge erforderlich.

2.3 Wartungsaufwand

Wartungsaufwand sind diejenigen Resourcen, die für die Verbesserung und Erweiterung des Systems benötigt werden.

Das Ausmaß des Aufwands für die Verbesserung und Erweiterung einer Liste oder eines Regelsystems unterscheidet sich ebenfalls drastisch. Während man eine Liste und die darin enthaltenen Strukturen sehr gut beschreiben und ihre Konsistenz und Qualität asymptotisch an eine Fehlerrate von 0% heranführen kann, ist die Veränderung eines Regelsystems vor allem kognitiv hoch aufwendig. Ebenfalls ist die Beschreibung der Funktionsweise und Vorgehensweise der einzelnen Regeln deshalb kompliziert, weil einfache generelle Regeln alleine wegen der unsystematischen Anteile der Orthographie nicht ausreichen und z.T. Winkelzüge - d.h. Regeln, die kaum in einer linguistischen Theorie ihren Platz finden würden - erforderlich sind.

Der Aufwand für die Aufrechterhaltung eines Lexikons ist recht gering, vorausgesetzt, die Dokumentation des Lexikons ist genau und verständlich. Der Aufwand für die Aufrechterhaltung und Anpassung eines Systems ist in jedem Falle größer, da das Wissen über die Systemeigenschaften weniger leicht kodierbar ist und ein Fortschreiten der technologischen Entwicklung die Veralterung und Unbrauchbarkeit der Systemumgebung unabwendbar macht.

Abschließend kann also gesagt werden, daß die Erstellung eines Lexikons in jedem Falle unabdingbar ist. Im nächsten Abschnitt werden die Möglichkeiten der Erstellung eines Aussprachelexikons erörtert.

3 Manuelle vs. automatische Transkription

Es wurde bereits festgestellt, daß es unmöglich ist, ein System zu erstellen, das mit einer Fehlerrate von 0% arbeitet. Dies ist allein schon deshalb undenkbar, als jedes (Test-)Lexikon ebenfalls Fehler aufweist. Menschliche Transkribenten hingegen haben ein verläßlicheres Wissen um die Aussprache von Wörtern. Leider jedoch ist menschliche Arbeitsleistung teuer und unkonstant, weil Faktoren wie Konzentration und Motivation starken Einfluß auf die Qualität der Arbeit haben, somit das tatsächlich bessere Wissen der Menschen keine Garantie für die korrekte Wiedergabe desselben ist. Darüberhinaus ist das Auffinden von von Menschen gemachten Fehlern überaus schwer, wenn die Bandbreite der zu erwartenden Fehler von Tippfehlern über Dialektabweichungen bis zu Nichtwissen reicht. Demgegenüber sind die von Systemen produzierten Fehler systematisch, dadurch leichter zu finden und zu korrigen.

Optimal ist deshalb jede Methode, die den Einsatz menschlicher Arbeit aus Geld-, Zeit und Qualitätsgründen minimiert und auf die Kontrolle von von Systemen produzierten Transkriptionen und die Wartung der Systeme beschränkt. Zusätzlich ist die Anwendung eines Strukturparsers, der die Wohlgeformtheit der automatisch generierten und von Hand korrigierten Einträge überprüft, anzuraten.

4 Automatische Transkriptionsmethoden

An automatischen Transkriptionssystemen lassen sich folgende Architekturen unterscheiden:

- regelbasiert: Es werden Ersetzungsregeln, die Buchstaben oder Lauten in definierten Kontexten andere Eigenschaften zuweisen oder sie ersetzen, formuliert.
- morphologiebasiert: Als Morpheme identifizierte Substrings werden in ihre jeweilige lautliche Entsprechung umgewandelt.
- neuronal: Mit Hilfe einer Datenbasis lernt ein neuronales Netz allgemeine Entsprechungen zwischen Buchstaben und Lauten.
- statistisch: Alle lautlichen Entsprechungen eines Buchstabens in allen in einer Datenbasis vorkommenden Kontexten werden in einer Datenbasis gespeichert.

Die Charakteristika der hier exemplarisch beschriebenen Systemarchitekturen werden in Tabelle 1 gezeigt.

Aspekt	regelbasiert	morphologiebasiert	neuronal	statistisch
Vorgabe	linguistisches Wissen	linguistisches Wissen	Datenbasis	Datenbasis
Arbeitsaufwand	Erstellen von Regeln	Erstellen von Lexikon und Regeln	Programmieren des Netzes	Programmieren der Datenanalyse
Speicheraufwand	gering 0,07 MB	mittel 0,21 MB	mittel 0,59 MB	sehr hoch 2,59 MB
Universalität	nein	nein	ja	ja
Operationseinheiten	bel. Fokus und Kontext	bel. Fokus kein Kontext	fester Fokus und Kontext	fester Fokus und Kontext
Ausgabeverhalten	vorhersagbar	vorhersagbar	nicht vorhersagbar	berechenbar
Verbesserbarkeit	konkrete Fälle	konkrete Fälle	allgemeine Leistung	allgemeine Leistung

Tabelle 1: Eigenschaften verschiedener Transkriptionssysteme

Unmittelbar einsichtig ist, daß jedes System Fehler aufweist. Will man sich aber tatsächlich ausschließlich auf die Verwendung von automatischen - im Gegensatz zu manuellen - Transkriptionsverfahren beschränken, etwa, weil die Anzahl der umzusetzenden Einträge zu groß und die Auftretenshäufigkeit einiger

Einträge zu gering ist, müssen besondere Verfahren eingesetzt werden. Die Qualität der jeweiligen Systeme läßt sich zwar in einer Fehlerrate ausdrücken. Für die jeweiligen Systeme schwer zu transkribierende Einträge bzw. fehlerhaft umgesetzte Einträge lassen sich aber ohne ein Lexikon nur zu einem gewissen Grad - etwa durch einen Strukturparser - abfangen. Eine Methode, mit unsicheren Ausgaben trotzdem eine sicherere Qualitätsangabe zu erreichen ist die folgende:

- Die Einträge eines vorhandenen Aussprachelexikons werden von jedem der verfügbaren Systeme umgesetzt.
- Die Fehlerraten für die einzelnen Systeme werden errechnet.
- Die Fehlerraten für alle Transkriptionen, die übereinstimmend von zwei, drei, ..., allen Systemen produziert werden, werden errechnet.
- Der letzte Schritt wird durch zusätzliche Eingrenzung unter Einbeziehung weiterer Informationen wie Wortart, Kasus etc. verfeinert.
- Es wird eine Rangliste der Verläßlichkeit der Transkriptionen einzelner und kombinierter Umsetzungen aufgestellt.
- Alle weiteren zu transkribierenden Einträge werden von allen Systemen umgesetzt.
- Die jeweils nach der oben berechneten Rangliste besten Transkriptionen werden ausgewählt und mit der Information ihrer Umsetzungsherkunft als Qualitätsangabe versehen.

5 Informationseinheiten in Aussprachelexika

Welche Informationseinheiten soll ein Aussprachelexikon bereitstellen? Die Angabe der Laute ist die augenscheinlich wichtigste Informationseinheit. Weitere Kandidaten sind: Silbengrenzen, Betonungsebenen (Haupt-, Nebenakzent, unbetont), Morphemgrenzen, Grundform, Wortart, Modus, Numerus, Kasus, Tempus und Häufigkeit. Die Angabe all dieser Informationen ist für einen besseren und gezielteren Zugriff bzw. Veränderungen nützlich. Integriert man morphologische Grenzen und die genannten phonetischen Informationen in die Transkription und verweist zusätzlich auf die anderen genannten formalen Aspekte, erhöht man die Brauchbarkeit des Lexikons für unterschiedlichste Untersuchungen. Ob zusätzliche semantische Informationen (semantische Felder, Bedeutungsrelationen, semantische Merkmale etc.) für phonetische Fragestellungen erforderlich sind, bleibt zu klären.

Alternativ lassen sich verschiedene Sortierungen des Lexikons denken: alphabetisch (nach der orthographischen oder phonetischen Information), rückläufig

(dito), morphologisch, häufigkeitsbasiert usw. Auch an diesen Beispielen zeigt sich nochmals die große Überlegenheit einer Datenbasis gegenüber eines Online-Systems: Die Daten sind permanent verfügbar und können so beliebig manipuliert werden.

6 Standardisierungsebenen

Unabhängig davon, ob Menschen oder Maschinen Wörter transkribieren: Es gibt verschiedene Ebenen, für die einheitliche Standards festgelegt werden müssen. Diese Ebenen sind die folgenden:

- die abzubildende Aussprachevarietät
- Bedingungen und Arten von Aussprachealternativen
- Elemente des Lautinventars
- Kodifizierung der möglichen Silbenübergänge
- zulässige Lautkombinationen
- Grammatik von Intonationskonturen
- Konsistenz der einheitlichen Transkription lautlich oder morphologisch ähnlicher Strukturen

Von der Präzision der Definition derartiger Standardisierungen und ihrer Einhaltung ist die Qualität des Lexikons ebenso abhängig wie von der Qualität der Einzeltranskriptionen.

7 Reihenfolge der Bearbeitung von Einträgen

Es empfiehlt sich nicht, die umzusetzenden und von Hand zu kontrollierenden Daten in der alphabetischen Reihenfolge ihres Auftretens zu bearbeiten. Ein Rechenbeispiel: Wenn die Einträge der Datenbasis gleichmäßig auf 20 verschiedene Anfangsbuchstaben verteilt sind, ist anzunehmen, daß nach der Umsetzung und Kontrolle der Einträge des Anfangsbuchstaben *a* 5% der Einträge und ebenfalls 5% der Nennungen transkribiert sind.[1] Andererseits könnte es vorkommen, daß die ersten 5% in einer nach Häufigkeit sortierten Liste der Daten bereits 70% der Nennungen abdecken. Deshalb sollte man häufigeren Einträgen größere Sorgfalt widmen und in der Reihenfolge der Transkription Priorität einräumen, da sie am häufigsten benutzt werden und dementsprechend wichtig sind.

[1] Wenn ein Wort in einer Datenbasis z.B. 20 mal auftritt, handelt es sich um einen Eintrag mit 20 Nennungen.