Sprachen auf dem WWW*

Stefan Langer

Zusammenfassung

Mit einem Sprachenidentifizierer für 45 Sprachen wurden die prozentualen Anteile der verschiedenen Sprachen an allen Webseiten durch eine repräsentative Analyse des Indexes von AllTheWeb vom Sommer 1999 ermittelt. Zunächst wird die Funktionsweise und die Performanz des verwendeten Sprachenidentifizierers kurz dargestellt Im Hauptteil werden die Ergebnisse der Index-Auswertung präsentiert. Die Statistiken zeigen sowohl die prozentualen Anteile der verschiedenen Sprachen an der Gesamtheit aller Index-Seiten als auch die Zahl der Webseiten pro Sprecher weltweit.

8.1. Einführung

Ziel dieser Publikation ist die Darstellung des prozentualen Anteils verschiedener Sprachen an Internetwebseiten. Grundlage ist eine Analyse des Indexes von AllTheWeb vom Sommer 1999. Mit Hilfe eines Sprachenidentifizierers für 45 Sprachen wurden die prozentualen Anteile von Webseiten in den verschiedenen Sprachen ermittelt. Vorab werden zur Einschätzung der Ergebnisse die Funktionsweise und Leistungsmerkmale des verwendeten Sprachenidentifizierers dargestellt. Der Schwerpunkt der Darstellung liegt dann auf der Darstellung der Ergebnisse unter zwei Gesichtspunkten: der prozentualen Häufigkeit der einzelnen Sprachen an den analysierten Webseiten und die Zahl der Webseiten in Relation zur Zahl der Sprecher einer Sprache. Letzere Zahl gibt einen Eindruck von der Durchdringung einzelner Sprachgemeinschaften durch das Internet.

8.2. Sprachenidentifizierung

Ein Sprachenidentifizierer (hier wird nicht der anderswo gelegentlich auftauchende Begriff Sprachenerkenner verwendet, um jede Verwechslungsmöglichkeit mit Programmen zur Spracherkennung gesprochener Sprache auszuschließen) ist Software zur automatischen Erkennung der Sprache(n) eines elektronischen Textdokuments. Sprachenidentifizierer werden v. a. in Internetsuchmaschinen und Textarchiven eingesetzt.

Alle großen internationalen WWW-Suchmaschinen verfügen inzwischen über die Möglichkeit, nur Dokumente anzuzeigen, die in einer oder mehreren präselektierten Sprachen verfasst sind. Die Zahl der unterstützten Sprachen ist dabei unterschiedlich hoch (vgl. Tab. 8.1).

Die Klassifizierung der Dokumente für alle Internet-Suchmaschinen erfolgt dabei nicht oder zumindest nicht überwiegend durch HTML-Metainformation im Dokument (z. B. Sprachen-

* Erschienen in: *Proceedings der GLDV-Frühjahrstagung 2001*, Henning Lobin (Hrsg.), Universität Gießen, 28.–30. März 2001, Seite 85–91. `http://www.uni-giessen.de/fb09/ascl/gldv2001/`

Google	25
Excite	11
Altavista	19
AllTheWeb	46

Tabelle 8.1.: Von populären Suchmaschinen unterstützte Sprachen (Stand: Januar 2001)

tags), da diese Informationen nur in wenigen Fällen gegeben werden und zudem sehr unzuverlässig sind, sondern durch eine Analyse des Dokumententextes mit einem Sprachenidentifizierer.

Eine weitere Anwendungsdomäne für Sprachenidentifizierer sind multilinguale Textarchive und das Intranet multinationaler Firmen. In der linguistischen Forschung wurden Sprachenidentifizierer zur automatischen Erstellung von Korpora aus dem WWW (Rehm, 2001) oder anderen multilingualen Textquellen eingesetzt (Cowie et al., 1998).

Für die automatische Erkennung der Sprache eines elektronischen Dokuments lassen sich zwei grundsätzlich verschiedene Techniken unterscheiden (für einen Performanzvergleich s. Grefenstette, 1995).

- Wortbasierte Sprachenerkennung. Die im Dokument gefundenen Wörter werden mit Frequenzlisten von Wortformen der unterstützten Sprachen abgeglichen und die Treffer gezählt. Die Sprache mit der besten Trefferquote, die sich über einem Schwellenwert befinden muss, gilt als identifiziert. Kaum anwendbar ist diese Technik auf Dokumente in Sprachen, die Wortgrenzen nicht systematisch durch Leer- oder Satzzeichen im Schriftbild markieren, wie etwa Japanisch und Chinesisch. Ein Vorteil dieser Technik ist, dass Sie sehr zuverlässig arbeitet, da die verwendeten Wortlisten manuell kontrolliert werden können. Fehlerquellen in den Daten können leicht geortet und beseitigt werden. Die notwendige Größe der Wortlisten hängt dabei eng mit der Länge der zu klassifizierenden Dokumente zusammen. Schon mit Hilfe einer Liste der häufigsten Wörter lassen sich längere Dokument sehr zuverlässig klassifizieren (Grefenstette, 1995) – je länger die Wortlisten werden, desto eher lassen sich auch kurze Dokumente einordnen, da die Trefferwahrscheinlichkeit erhöht wird.

- N-Gramm-basierte Sprachenerkennung (z. B. Cavnar und Trenkle, 1994). Hierbei werden N-Gramme von Buchstaben oder Bytes verwendet und mit den Statistiken über einen Trainingskorpus für jede Sprache abgeglichen. Es werden meist Statistiken über Trigramme verwendet. Diese Technik ist universell verwendbar, allerdings problematisch für sehr kurze Dokumente. Außerdem sind Fehlerquellen schwierig zu eliminieren, da die N-Gramm-Statistiken kaum manuell zu kontrollieren sind.

Alle Techniken zur Sprachenidentifizierung setzen eine Parser (d. h. einen Extraktor für natürlichsprachliche Textabschnitte) für den analysierten Dokumententyp voraus. Der für unsere Untersuchungen verwendete Sprachenidentifizierer ist ein hybrider Sprachenidentifizierer, d. h. er verwendet beide Techniken. Zunächst wird eine wortbasierter Algorithmus verwendet. Dazu wurden Wortformlisten mit Frequenzangaben für 41 Sprachen erstellt. Führt dieser zu keinem Ergebnis, wird überprüft, ob er über ein bigrammbasiertes Verfahren eine der vier Sprachen er-

kannt werden kann, die Wortgrenzen nicht systematisch markieren (in der verwendeten Version Japanisch, Chinesisch, Koreanisch und Thai).

Der Sprachenidentifizierer identifiziert gleichzeitig mit der Sprache eines Dokuments auch dessen Zeichensatzkodierung – die Information über die verwendeten Kodierung für die verschiedenen Sprachen wurde aber für die hier vorgelegten Statistiken ignoriert.

Für Dokumente mit mehr als 30 Zeichen Text arbeitete der Identifizierer auf einem Testkorpus mit 100 Dokumenten in jeder unterstützen Sprache mit einem Recall von ca. 96% und einer Precision von 100%. Letzterer Wert muss für die Anwendung auf extrem distribuierten Dokumentensammlung wie dem WWW sehr hoch sein, da eine Fehlklassifizierung von auch nur wenigen Dokumenten einer häufigen Sprachen zu extremen Veränderungen der Zahl der gefundenen Dokumente für eine fälschlicherweise erkannte seltene Sprache haben könnte. Da die Qualität der Klassifizierung für die vorgenommene Untersuchung ausreichend ist, erübrigte sich ein systematischer Vergleich mit einem rein N-Gramm-basierten Sprachenerkenner.

8.3. Untersuchungsprämissen

Mit dem vorgestellten Sprachenidentifizierer wurde die primäre Sprache von 3 Millionen Webseiten mit mindestens 30 Zeichen ermittelt, die nach dem Zufallsprinzip aus dem vollständigen Index von AllTheWeb[1] vom Juli 1999 ausgewählt wurde. Der gesamte Index umfasste ca. 300 Millionen Seiten. Web-Indexe können insofern als recht repräsentativ für das gesamte WWW gelten, als sie tatsächlich zugängliche und verlinkte Seiten präferiert indizieren – d. h. sie repräsentieren das WWW in der Form, wie es sich auch den Internetnutzern darstellt (die zudem häufig über Suchmaschinen auf das Netz zugreifen und damit ausschließlich auf indizierte Seiten Zugriff haben). Das in O'Neill et al. (1997) vorgestellte aleatorische Verfahren durch die Generierung zufälliger Server-IP-Nummern dagegen erfasst auch solche Web-Sites, die für WWW-Benutzer nicht auffindbar sind.

8.4. Ergebnisse

Die Tabellen stellen die Ergebnisse der Untersuchung dar. Für 94% aller Seiten wurde eine Sprache eindeutig identifiziert. Die Zahl der nicht-klassifizierten Seiten lag bei 6%. Eine Stichprobenanalyse ergab folgende Gründe dafür, dass eine Seite nicht klassifiziert werden konnte:

- das Dokument war auch manuell keiner Sprache eindeutig zuzuordnen (etwa Dateilisting, Datenmüll, Namenslisten)
- das Dokument enthielt nur wenige, niederfrequente Wörter einer unterstützten Sprache
- das Dokument war in einer nicht-unterstützten Sprache verfasst (eher selten, die Analyse deckt anscheinend alle quantitativ wichtigen Sprachen auf dem WWW ab)
- die Seite war mehrsprachig

Das Ergebnis der quantitative Analyse wird in Tab. 8.2 zusammengefasst. Sie stellt den prozentualen Anteil der verschiedenen Sprachen auf dem Netz dar. Nicht verwunderlich ist die Dominanz des Englischen – 2/3 der Webseiten sind in englischer Sprache verfasst. Auch die anderen

[1] `http://www.AllTheWeb.com`

Pos.	Sprache	%	Pos.	Sprache	%
1	Englisch	65,0	24	Griechisch	0,083
2	Japanisch	6,0	25	Indonesisch, Malai	0,082
3	Deutsch	5,0	26	Slowenisch	0,058
4	Französisch	3,1	27	Ukrainisch	0,052
5	Chinesisch	2,3	28	Hebräisch	0,052
6	Spanisch	2,3	29	Isländisch	0,050
7	Italienisch	1,7	30	Estnisch	0,045
8	Russisch	1,6	31	Rumänisch	0,043
9	Portugiesisch	1,1	32	Bulgarisch	0,022
10	Schwedisch	1,0	33	Vietnamesisch	0,020
11	Koreanisch	0,97	34	Litauisch	0,019
12	Niederländisch	0,94	35	Lettisch	0,018
13	Tschechisch	0,59	36	Baskisch	0,014
14	Finnisch	0,55	37	Arabisch	0,012
15	Norwegisch (Bokmal und Nynorsk)	0,42	38	Latein	0,0076
			39	Afrikaans	0,0073
16	Polnisch	0,41	40	Galizisch	0,007
17	Dänisch	0,36	41	Walisisch	0,0048
18	Ungarisch	0,24	42	Weißrussisch	0,0018
19	Katalanisch	0,16	43	Färöisch	0,0015
20	Slowakisch	0,11	44	Westfriesisch	0,0004
21	Thai	0,11			
22	Türkisch	0,10		Nicht klassifiziert	6
23	Kroatisch	0,086			

Tabelle 8.2.: Sprache in % aller Webseiten (Grundlage: Index AllTheWeb, Sommer 1999)

Pos.	Sprache	S/Spr.	Pos.	Sprache	S/Spr.
1	Isländisch	6,2	24	Russisch	0,32
2	Schwedisch	3,5	25	Ukrainisch	0,32
3	Finnisch	2,7	26	Polnisch	0,28
4	Norwegisch	2,6	27	Kroatisch	0,26
5	Dänisch	2,2	28	Spanisch	0,25
6	Japanisch	1,5	29	Walisisch	0,25
7	Deutsch	1,5	30	Portugiesisch	0,21
8	Tschechisch	1,5	31	Griechisch	0,20
9	Niederländisch	1,4	32	Thai	0,16
10	Englisch	1,4	33	Litauisch	0,15
11	Estnisch	1,3	34	Chinesisch	0,07
12	Färöisch	1,2	35	Bulgarisch	0,07
13	Slowenisch	0,8	36	Türkisch	0,05
14	Italienisch	0,8	37	Rumänisch	0,05
15	Katalanisch	0,8	38	Afrikaans	0,037
16	Baskisch	0,7	39	Weißrussisch	0,027
17	Slowakisch	0,5	40	Westfriesisch	0,017
18	Galizisch	0,5	41	Indonesisch, Malai	0,015
19	Koreanisch	0,5	42	Vietnamesisch	0,01
20	Ungarisch	0,5	43	Arabisch	0,002
21	Französisch	0,4			
22	Hebräisch	0,4	44	Latein	keine
23	Lettisch	0,36			Sprecher

Tabelle 8.3.: Webseiten/Sprecher (zugrunde liegt eine Zahl von 300 Mio. Webseiten – gerundete Zahl der Webseiten im Index von AllTheWeb, Sommer 1999; zugrundeliegenden Sprecherzahlen sind Schätzungen nach Grimes, 1997)

Spitzenplätze sind wenig überraschend – es handelt sich hier entweder um die Sprachen hochindustrialisierter Länder mit hoher Bevölkerungszahl (Japanisch, Deutsch) oder um Sprachen von Ländern mit eine sehr hohen Bevölkerungszahl und einem gewissen Industrialisierungsgrad (Chinesisch). Die letzten Listenplätze nehmen Sprachen mit sehr wenigen Sprechern (Färöisch), weniger verbreitete Minderheitensprachen (Galizisch, Westfriesisch) und Sprachen aus Ländern mit einem geringen Industrialisierungsgrad ein (Weißrussisch).

Tab. 8.3 zeigt das Verhältnis zwischen der hochgerechneten Zahl der Webseiten im gesamten Index zu den Sprecherzahlen der unterstützten Sprachen, und gibt damit einen Hinweis auf die Internet-Affinatät der verschiedenen Sprachgemeinschaften. Der Analyse liegt die Zahl von 300 Millionen Webseiten und die Sprecherzahlen in Grimes (1997) zugrunde – die notwendigerweise zum Teil nur geschätzte Zahlen sind. In dieser Statistik liegen die skandinavischen Länder weit vorn – das Englische nimmt (aufgrund der zahlreichen Sprecher in nicht-industrialiserten Ländern) keinen Spitzenplatz mehr ein. Auffällig ist bei beiden Statistiken die geringe Relevanz des Arabischen – der prozentuale Anteil arabischer Seiten an der Gesamtheit der Dokument ist geringer als der Anteil baskischer Webseiten und in Bezug auf die Sprecherzahl liegt das Arabische weit abgeschlagen auf dem letzten Platz – neben soziolinguistischen und wirtschaftlichen Gründen spielt hier sicher eine Rolle, dass Webbrowser lange die arabische Schrift aufgrund Ihrer komplexen Graphie nicht unterstützten.

8.5. Verwandte Arbeiten

Lavoie und O'Neill (1999) stellen eine Analyse von 1 257 (1998) und 2 229 (1999) Web-Sites vor. Die verwendeten Adressen wurden durch ein Zufallsverfahren für die Generierung von IP-Nummern von Web-Servern ausgewählt. Für diese wurde das Land manuell (über angegebene Adressen) und die Sprache teilweise manuell und teilweise automatisch ermittelt. Die Ergebnissen für 1999 unterscheiden sich teilweise von den von uns vorgestellten – allerdings lassen sich die Zahlen aufgrund des unterschiedlichen Stichprobenverfahrens, des äußerst verschiedenen Umfangs der Untersuchung und der anderen unterschiedlichen Prämissen nur schwer vergleichen.

Grefenstette und Nioche (2000) ermitteln durch die Ermittlung der Frequenz häufiger Wörter die anzunehmende Wortzahl von 32 Sprachen im Index von AltaVista. Sie beschränken sich dabei auf Sprachen mit lateinischem Alphabet. Für die Sprachen, die in vorliegender Untersuchung ebenfalls erfasst sind, kommen sie zu quantitativ vergleichbaren Ergebnissen – d. h. die prozentualen Anteile der abgedeckten Sprachen stimmen ungefähr mit unseren Werten überein, obwohl nicht die Zahl der Webseiten einer Sprache, sondern die Gesamtwortzahl im Index hochgerechnet wird.

8.6. Offene Fragen

Einige Fragen sollen in einer weiteren Untersuchung geklärt werden:

- Wie ist die Entwicklung der Sprachen auf dem WWW? Welche Sprachen wachsen prozentual, welche fallen? Grefenstette und Nioche (2000) zeigen mit drei Stichproben für acht Sprachen die Entwicklung der geschätzten Wortzahlen im gesamten Index von AltaVista von 1996–2000 auf und belegen, dass der prozentuale Anteil des Englischen – wie zu erwarten – rückläufig ist. Ihre Liste beschränkt sich aber auf (ursprünglich) europäische Sprachen.

- Welche weiteren Sprachen sind relevant? Der vorgestellte Sprachenerkenner strebt eine weitgehend vollständige Erfassung aller Sprachen auf dem Netz an. Zu diesem Zweck müssten nicht erkannte Dokumente ausgewertet und die Wörterbücher noch um einige Sprachen erweitert werden. Bereits ermittelte Kandidaten sind Bretonisch, irisches Gälisch, Farsi, Esperanto; hinzu kommen sicher noch Sprachen Afrikas, Asiens und Minderheitensprachen anderer Kontinente, die bisher eine geringe quantitative Relevanz auf dem Netz haben.

Literaturverzeichnis

Cavnar, W. B. und Trenkle, J. M. (1994): "N-Gram-Based Text Categorization". In: *Symposion On Document Analysis and Information Retrieval*. University of Nevada, Las Vegas, S. 161–176.

Cowie, J.; Ludovik, E. und Zacharski, R. (1998): "An Autonomous, Web-based, Multilingual Corpus Collection Tool". In: *Proceedings of the International Conference on Natural Language Processing and Industrial Applications*. Moncton, S. 142–148. Online verfügbar:FIXME: URL einfuegen!

Grefenstette, G. (1995): "Comparing two language identification schemes". In: *Proceedings of the 3rd International Conference on the Statistical Analysis of Textual Data (JADT '95)*.

Grefenstette, G. und Nioche, J. (2000): "Estimation of English and non-English Language Use on the WWW". In: *Proceedings of RIAO'2000: Content-Based Multimedia Information Access*. Paris, S. 237–246. Online verfügbar: `http://de.arXiv.org/abs/cs.CL/0006032`.

Grimes, B. F. (Herausgeber) (1997): *Ethnologue: Languages of the World*. Dallas: SIL Publications. Siehe auch `http://www.sil.org/ethnologue/`.

Lavoie, B. F. und O'Neill, E. T. (1999): "How "World Wide" is the Web? Trends in the Internationalization of Web Sites". In: *Annual Review of OCLC Research 1999*, Dublin: OCLC Online Computer Library Center. Online verfügbar: `http://www.oclc.org`.

O'Neill, E. T.; McClain, P. D. und Lavoie, B. F. (1997): "A Methodology for Sampling the World Wide Web". In: *Annual Review of OCLC Research 1997*, Dublin: OCLC Online Computer Library Center. Online verfügbar: `http://www.oclc.org`.

Rehm, G. (2001): "`korpus.html` – Zur Sammlung, Datenbank-basierten Erfassung, Annotation und Auswertung von HTML-Dokumenten". In: *Proceedings der Frühjahrstagung der Gesellschaft für Linguistische Datenverarbeitung*. Justus-Liebig-Universität Gießen. In diesem Band.