LEKTOR - ein Programm zur Lesetext-Evaluation

関西外国語大学 研究論集 第87号（2008年3月）
Journal of Inquiry and Research, No.87 (March 2008)

LEKTOR - ein Programm zur Lesetext-Evaluation

Ruediger Schmidt

Abstract

Die vorliegende Arbeit stellt das von mir "LEKTOR" genannte, für das Microsoft Windows Betriebssystem entwickelte Programm vor. Es ermöglicht einer im Bereich "Deutsch als Fremdsprache" (DaF) tätigen Lehrperson, Texte auf ihre Leseschwierigkeit für Lerner zu überprüfen. Dabei gestattet das Programm folgende Dinge:

- den Schwierigkeitsgrad eines Textes zu bestimmen
- eine Wortliste der Wörter auszugeben, die den festgelegten Schwierigkeitsgrad übersteigen
- allgemeine statistische Informationen zum Text zu erhalten wie Wortzahl, Satzzahl, etc.
- eine grobe syntaktische Analyse der Sätze auszugeben

Um all das zu ermöglichen, wurde ein Grundwortschatz von ca. 1600 Wörtern gewichtet/gradiert und in 10 Stufen eingeteilt. Die Wörter des Textes werden analysiert, mit den entsprechenden Lexikoneintragungen verglichen und je nach Vorgabe eine Liste erstellt und ausgegeben.
Für die syntaktische Analyse wurde kein formalisiertes Sprachmodell verwendet, weil eine solche Analyse nicht das Hauptanliegen für die gestellte Aufgabe war. Stattdessen wurden Methoden, wie sie bei PoS-Taggern angewendet werden, sowie eigene heuristische Ansätze eingesetzt.

Keywords: Daf, maschinelle Textanalyse, Wortschatz, graded reading

A Einleitung: Motivation zur Erstellung des Programms

Unbestritten ist zwar, dass man eine Sprache nicht nur unter kommunikativen Gesichtspunkten lehren kann, bei denen Hören und Sprechen im Vordergrund stehen, dass zu den vier Grundfertigkeiten eben auch Lesen und Schreiben gehören.
Trotzdem macht man immer wieder die Erfahrung, dass sich die Studenten schwer tun, vor al-

lem beim Schreiben, aber auch beim lesenden Verstehen.
Dieses Manko liegt zum Teil darin begründet, dass - je nach Universität - diese Fertigkeiten zu wenig geübt werden. Sehr schnell werden sie mit schwierigen Texten konfrontiert, die sie ohne Hilfe des Lehrers gar nicht verstehen könnten. Was ihnen letztlich fehlt ist die vielfältige **Erfahrung der deutschen Sprache**, die ihre noch auf wackeligen Beinen stehenden Grundkenntnisse zu festigen vermag. Nun, eine Möglichkeit, diese Erfahrung zu erwerben, - im Ausland vielleicht sogar die beste, weil man hier ja nicht jeden Tag Deutsch hört und spricht - besteht **im Lesen**. Durch Lesen gewinnt der Lerner Einblicke in das **Wirken** der deutschen Sprache, Einsichten, die bewusst/unbewusst und intuitiv stattfinden und sich auch in Aha-Erlebnissen äußern: so werden die Wörter verwendet, so ist der Satz aufgebaut, so sind die Sätze miteinander verknüpft usw.
Waring (2003) hat für das Englische darauf hingewiesen, dass man durch ein dem Lerner angepasstes Lesen nicht unbedingt neuen Wortschatz erwirbt, dass aber ein Gewinn sprachlicher Erfahrung sicher ist:

> "[...] that ultimately learners do not learn a lot of new words from graded reading, but in fact graded reading helps to deepen and consolidate *already known* language."[1]

Wie im Englischen wäre also auch für das Deutsche ein "graded reading" erforderlich, ein der Lernstufe und dem Niveau des Lerners angepasstes Lesen. Während aber für das Englische sogenannte "Easy Readers" schon ab einem Wortschatz von ca. 150 Wörtern erhältlich sind, setzen die meisten für das Deutsche erhältliche Lesehefte 600-700 Wörter voraus, definitiv zu viel.
Außerdem ist die Zahl der fürs Deutsche erhältlichen Reader - verglichen mit den englischen - enttäuschend gering. Eine Lösung wäre, Texte selber zu schreiben oder aber bereits vorhandene Texte zu vereinfachen, aber auch auf diese Weise ist man nicht von dem Problem befreit, die **Leseschwierigkeit** irgendwie sinnvoll, reproduzierbar und (in Grenzen) objektiv festlegen zu können. Diese Kriterien legen nahe, hier die Hilfe des Computers in Anspruch zu nehmen.

B Aufbau des Lexikons und des Programms

1. Allgemeine Überlegungen

Es liegt auf der Hand, dass Programm und Lexikon aufeinander abgestimmt sein müssen, d.h.

alle Informationen, die das Programm nicht erbringen kann oder will, müssen ins Lexikon aufgenommen werden, und umgekehrt.
Welche **Mindestanforderungen** sind dabei an Lexikon und Programm zu stellen? Bevor man diese Frage beantworten kann, muss zuerst der grundlegende Aufbau des Lexikons entschieden sein.
Dazu gehört auch die Anzahl der Wörter, die ins Lexikon aufgenommen werden soll. Wenn wir davon ausgehen, dass die zu bearbeitenden Texte im 2. bzw. 3. Studienjahr gelesen werden, ist ein Grundwortschatz von ca. 2000 Worten sicherlich ausreichend.
Für das Lexikon gibt es prinzipiell 2 Möglichkeiten: entweder Vollformen-Lexikon oder nur Grundformen-Lexikon. Beide haben ihre Vor- und Nachteile.
Das Vollformen-Lexikon ist sicherlich im Ganzen effizienter, und in Anbetracht der heutigen Ausstattung der PCs an Speichermedien bei der geringen Wortzahl auch nicht zu umfangreich. Da das Eintippen aller Formen per Hand zu mühsam ist, braucht man allerdings einen Generator, der aus den Grundformen der flektierbaren Wortarten (vor allem Nomen, Verben, Adjektive) sämtliche anderen Formen generiert.
Das Grundformen-Lexikon ist einfach zu erstellen und auch zu warten. Da es aber nicht alle Formen enthält, ist innerhalb des Programms eine Lemmatisierung erforderlich. Dadurch gestaltet sich bei Änderungen die Wartung des Programms etwas aufwendiger. Andererseits kann das Grundformen-Lexikon vollständig in den Speicher eingelesen werden, entsprechend schnell ist das Aufsuchen eines Wortes.
Da ich auf keinerlei Hilfe außer meiner selbst zurückgreifen konnte, habe ich mich aus zeitlichen Gründen für das Grundformen-Lexikon entschieden.

Mindestanforderungen an das Lexikon:

- der ausgewählte Wortschatz muss für jedes Wort eine Kennung zum Schwierigkeitsgrad enthalten
- der Wortschatz soll eine Klassifizierung der Wortart enthalten

Für das Programm gilt:

- es muss einen als Textdatei vorliegenden (im Prinzip beliebig langen) Text in Sätze und diese wiederum in Wörter zerlegen

- es muss diese Wörter im Lexikon nachschlagen (vorhanden / nicht vorhanden), die Grundform aus der Oberfächenform ableiten sowie den Wortschatz-Grad und die Wortart notieren
- es muss eine Liste ausgeben aller Wörter, die über dem gewählten Wortschatz-Grad liegen bzw. im Lexikon nicht enthalten sind
- es muss anhand der Ausgabe im Vergleich zur Wortzahl des Textes den Schwierigkeitsgrad ermitteln

2. Gestaltung des Lexikons

Zunächst einmal gilt es festzulegen, wie der Wortschatz aufbereitet werden soll, welche Wörter man aufnimmt und welche nicht, welchen Grad man ihnen zuweist und dergleichen mehr. In Anschluss daran muss man überlegen, in welche Wortklassen der so gewonnene Wortschatz eingeteilt wird und welche Kategorien/Merkmale syntaktischer, morphologischer und semantischer Art man in das Lexikon aufnimmt.

a) Wortschatz

Da der Umfang des Lexikons - seinem Zweck entsprechend - von vornherein limitiert ist, muss man der Auswahl der Wörter besondere Beachtung schenken. Man kann natürlich einfach die Wortliste des "Zertifikats Deutsch" zugrunde legen, aber damit hat man noch keine Rangfolge. Hier bietet es sich an, Häufigkeitslisten zu Rate zu ziehen, was ich auch getan habe. Allerdings ist die in solch einer Häufigkeitsliste abgebildete Rangfolge stark abhängig von dem Korpus, aus dem sie gewonnen wurde. So sind z.B. die an der Universität Leipzig erstellten Häufigkeitslisten[2] meines Erachtens von einem Deutsch geprägt, wie man sie in den Medien, sprich: Nachrichten findet. Das Wort "Prozent" rangiert in der Häufigkeit auf Platz 63, das Wort hat aber im DaF-Unterricht einen wesentlich geringeren Stellenwert. Solche Häufigkeitslisten darf man also nicht unbedacht übernehmen.

Am besten gefielen mir schließlich die Häufigkeitslisten[3], wie sie in Deutschland für den Grundschulbereich zusammengestellt wurden. Die decken immerhin die ersten 600 Wörter ab und haben gleichzeitig einen hohen Gebrauchswert. Darauf aufbauend habe ich dann sonstige Häufigkeits- und andere Wortschatzlisten durchgesehen, eine Auswahl daraus getroffen und nach eigenem Ermessen, aber unter Berücksichtigung der Wortschatzlisten der Universität

Leipzig eine Reihenfolge getroffen.
In der jetzigen Form enthält das Lexikon 1620 Wörter, in 10 Stufen aufgeteilt, so dass jede Stufe etwa 160 Wörter enthält.

Lex – Stufe	1	2	3	4	5	6	7	8	9	10
Wortzahl	160	320	480	640	800	960	1120	1280	1440	1600

Obwohl der Wortschatz damit festliegt, muss man noch überlegen, welche Informationen den Lemmata beigegeben werden soll. Erst dann kann das Lexikon erstellt werden. Welche Punkte hier zu bedenken sind, soll im Folgenden kurz erläutert werden.

a) Vollformen vs. Stammformen

Hilfsverben, Konjunktionen, Pronomina, Artikelwörter und Präpositionen wurden komplett als Vollformen ins Lexikon übernommen. Das sind geschlossene Listen, deren Bestand sich nicht erweitert und deren Übernahme als Vollform das Lexikon nicht belastet.

b) Wortklassen

Je nach den Leistungsvorgaben des Programms ist auch die Einteilung der Wörter in Wortklassen sehr unterschiedlich, vor allem, was Subklassifizierungen betrifft. Das ist sicher erforderlich, wenn eine genaue Syntaxanalyse angestrebt wird. Im vorliegenden Fall jedoch dient die zugegebenermaßen grobe Syntaxanalyse nur dem Zweck, unnötige Analysen bei der Lemmatisierung zu übergehen. Für unsere Anwendung benutzen wir folgende Wortklassen:
Nomen, Personalpronomen, Fragepronomen, Possessivpronomen, Verb, Hilfsverb, Adjektiv, Adverb, Artikelwörter (best./unbest. Artikel, Demonstrativpron.), Konjunktionen, Zahl sowie Unb (unbekannt) für Wörter, die keiner der vorstehenden Wortklassen zugeordnet werden konnten bzw. im Wörterbuch nicht enthalten sind.

c) Probleme in der Wortklasse ''Verb''

Ein heikler Punkt ist die Behandlung der Verben mit trennbarem Präfix wie z.B. abfahren: Der Zug fährt um 9:45 Uhr von Köln ab. Es empfiehlt sich hier, das Stammverb ''fahren'' als ''trennbarer Präfix ist möglich'' zu markieren und für diese Verbgruppe ein getrenntes Lexikon anzulegen. Dieses Feature ist in der gegenwärtigen Version noch nicht implementiert[4].

Die für das Lexikon zusammengestellten Lemmata liegen zunächst in einer Textdatei vor, die aber aus programmtechnischen Gründen in eine binäre Datei umgewandelt wurde. Einträge in der Text-Datei haben die Form "Lemma, Wortklasse":

Haus, Nom geh, Vrb mit, Präp

Zu jedem Eintrag im Lexikon wird außerdem ein Hash-Wert als Index errechnet und in eine Hash-Tabelle eingetragen, um die Suche nach dem betreffenden Wort effizient zu gestalten.

3. Programmfluss

Im Folgenden soll der Ablauf des Programms kurz skizziert werden.

1. Der Text wird eingelesen und satzweise in einer Array-Struktur **SI** abgespeichert
2. Ein Satz wird in Wörter aufgespaltet und in einer Array-Struktur **WI** abgespeichert
3. Die Wörter werden in einem ersten Durchgang im Lexikon gesucht und, falls gefunden, die Informationen dazu in die Struktur eingetragen
4. Erste Disambiguierung von Wörtern, die in zwei Wortklassen auftreten; Wörter, die nun eindeutig einer Wortklasse zugeordnet werden konnten und deren Oberflächenform im Lexikon enthalten ist, werden entsprechend markiert
5. Der Satz wird auf Phrasen untersucht und die bislang nicht gefundenen Wörter werden gezielt einer lexikalischen Analyse unterzogen; wenn gefunden, werden auch sie markiert
6. Sämtliche Informationen zu den Wörtern werden in **SI** eingetragen
7. Wenn das Textende noch nicht erreicht ist, springt das Programm zu 2. zurück
8. Textende erreicht, Ausgabe der Wortliste und sonstiger Infos

4. Programm-Einzelheiten

Wie auf Lexikonebene, so gibt es auch auf der Programmebene einige Hindernisse zu überwinden.

1. Satzende erkennen

Das Satzende ist erreicht, sobald ein Satzende-Zeichen (".!?") gelesen wird.
Allerdings tritt der Punkt "." auch in Abkürzungen usf. auf:

(1) z.B. (2) Dr. Müller (3) Herrn C. (4) am 3. Oktober (5) um 15.45 Uhr

Während man in diesem Fall mittels Abkürzungen-Lexikon "z.B." und "Dr." leicht erkennen kann, ist das in den nächsten Beispielen schon schwieriger:

(6) Ich traf Herrn C. in Berlin.
(7) Die Feier ist am 3. Oktober.
(8) Ich komme erst um 3. Lena kommt auch.

Dem Punkt als Satzende-Zeichen muss ein Leerzeichen folgen und der darauf folgende Buchstabe muss eine Majuskel sein - diese Bedingung ist zwar notwendig, aber nicht hinreichend: während (6) dadurch gelöst wird, versagt diese Regel in (7). In (8) aber weist der Punkt nach der Zahl "3" nicht auf eine Ordinalzahl, sondern auf ein reales Satzende.

2. Disambiguierung

Das gleiche Wort kann verschiedenen Wortklassen angehören:

(9) Das Buch gehört einem Kind. (Artikel)
(10) Da kann einem ja schlecht werden. (Indefinitpronomen)
(11) Einigen hat er Geld geschenkt. (Indefinitpronomen)
(12) Einigen konnten sie sich aber nicht. (Verb)

Diese Ambiguität aufzulösen kann mitunter sehr schwer sein. Im vorliegenden Programm werden (9), (10) und (11) richtig zugeordnet, (12) jedoch nicht. Das liegt aber zum einen daran, dass bei der Disambiguierung die vorhandenen syntaktischen Informationen noch nicht vollständig ausgewertet werden, zum andern an dem sehr eingeschränkten Tag-Set (bzw. Anzahl der Wortklassen).

3. Nomenanalyse

Die Nomenanalyse beginnt mit dem ersten Buchstaben eines Wortes. Ist er eine Majuskel, so wird dieses Wort tentativ als Nomen markiert. Da das erste Wort im Satz auch mit einem Großbuchstaben geschrieben wird, muss hier auf die Zugehörigkeit zu einer anderen Wortklasse geprüft werden.

Die Duden-Grammatik unterscheidet bei der Deklination 10 Typen, die auf unterschiedliche Kombination von 3 Singulartypen (S1-S3) und 5 Pluraltypen (P1-P5) aufbauen. Für die Analyse kann man das zunächst beiseite lassen, wichtig jedoch ist die Häufigkeit dieser 10 Typen. Die wichtigsten sind laut Duden Typ 1, 2 und 9, auf die ca. 90% der Substantive entfallen:

Typ 1:	das Jahr - des Jahres - die Jahre	30%
Typ 2:	das Muster - des Musters - die Muster	9%
Typ 9:	die Frau - der Frau - die Frauen	49%[5]

Um das Nomen von der Oberflächenform auf seine Stammform zurückzuführen, brauchen nur Plural- und Kasusmorpheme betrachtet zu werden:

Pluralmarker		
-x	(ohne Endung)	die Koffer
‘‘-x	(ohne Endung, aber Umlaut)	die Gärten
-e		die Berge
‘‘-e	(Endung -e + Umlaut)	die Hüte
-se		die Geheimnisse
-n	(niemals Umlaut!)	die Hosen
-en	(niemals Umlaut!)	die Frauen
‘‘-er	(Umlaut)	die Männer
-s	(niemals Umlaut!)	die Autos
-ien		die Indizien
-nen		die Lehrerinnen
Kasusmarker		
-s	(Genitiv)	des Bürgers

-es	(Genitiv)	des Jahres
-ns		des Namens
-ens		des Herzens
[-e]	(Dativ)	dem Jahre
-n	(Dativ Plural)	den Regeln

Sortiert man die Marker, so ergeben sich folgende vier Regeln, auf die die nicht sofort im Lexikon gefundenen Nomen überprüft werden:

1. [-UML]
2. [-UML]-[s][e]
3. [-UML]-[er, [i, n]e][n]
4. [[s]e, [e]n][s]

4. Adjektivanalyse

Werden Adjektive attributiv gebraucht, so stehen sie vor dem Nomen und werden dekliniert. Ihre Deklination hängt davon ab, ob und was für ein Artikelwort vor ihnen steht.
Eine zweite Flexionsform ist die Steigerung (Komparativ und Superlativ). Auch diese Formen können attributiv vor einem Nomen stehen. Aus Gründen der Einfachheit wurden Komparativ und Superlativ zusammen mit dem Positiv ins Lexikon aufgenommen. Bei Adjektiven, die auf -er oder -el enden, wurde das um -e- verkürzte Allomorph (z.B. teur- in "ein teurer Wein", ebenso dunkl-) der Einfachheit halber mit ins Lexikon aufgenommen. Da die Anzahl dieser Adjektive verschwindend gering ist, wird das Lexikon dadurch nur unbedeutend belastet.

Trotz der Komplexität der Kombination aus Artikelart, Genus, Numerus und Kasus gibt es nur fünf Flexionsendungen: -e, -em, -en, -er und -es. Da es auch hier genügt, zu einer gegebenen Oberflächenform die Stammform zu finden, gestaltet sich der Algorithmus sehr einfach:

1. [e]
2. [e[m, n, r, s]]

Ein spezielles Analyseproblem stellt der adverbial gebrauchte Superlativ da: Dieser Wein schmeckt mir am besten. Da dem Superlativ das Lemma "am" vorausgeht, wird im Falle eines als unbekannt eingestuften Wortes darauf geprüft und bei positivem Entscheid das Wort an die Adjektivanalyse übergeben.

5. Verbanalyse

Die Verbanalyse ist durch den flexionsbedingten Formenreichtum das aufwendigste Modul im Programm.

Hinsichtlich der Flexion unterscheiden wir bei den Verben drei Gruppen: schwache und starke Verben sowie als dritte Gruppe gemischt konjugierende Verben, die Eigenschaften beider vorhergehender Klassen besitzen.

Die finiten Verbformen sind morphologisch von folgenden Merkmalen abhängig:

- Person: 1., 2. und 3. Person
- Numerus: Singular und Plural
- Tempus: Präsens und Präteritum
- Modus: Indikativ, Konjunktiv und Imperativ

Da die anderen Tempora (Futur I+II, Perfekt, Plusquamperfekt) mittels Hilfsverben realisiert werden, brauchen wir sie bei der Verbanalyse nicht berücksichtigen.

Außerdem können Verben neben dem Infinitiv noch zwei Partizipien bilden:

- Partizip I (Partizip Präsens)
- Partizip II (Partizip Perfekt)

Weiterhin kann man Verben mit und ohne Präfix unterscheiden. Verben mit Präfix lassen sich in Verben mit trennbarem Präfix und solche mit untrennbarem Präfix einteilen. Da ich mich aus Zeitmangel für eine sehr einfache Lexikonstruktur entschieden habe, ist diese Gruppe nur unzureichend implementiert.

Die Hilfs- und Modalverben wurden komplett als Vollformen ins Lexikon aufgenommen und können hier also übergangen werden.

a) schwache Verben

Die schwachen Verben bilden ihre finiten Formen mittels der vier Flexionssuffixe -e, -st, -t und -en.

Diese vier Flexionssuffixe können sich unter bestimmten Bedingungen ändern:

- Bei Verben wie klingeln oder steuern ist das Flexionssuffix nicht -en, sondern -n.
- Endet der Verbstamm auf -t, -d, -m, -n, so steht in der 2. Person Singular und Plural sowie in der 3. Person Singular ein "-e-" vor dem Flexionssuffix.
- Endet der Verbstamm auf -s, -ss, -ß, -tz, -z, so wird das Fexionssuffix in der 2. Person Singular zu -t.
- Im Konjunktiv stehen statt -st und -t immer -est und -et.

Die schwachen Verben bilden das Partizip I mit dem Suffix -end, das Partizip II mit dem Präfix ge- und dem Suffix -t und den Infinitiv mit dem Suffix -n bzw. -en.

Somit müssen Verben, sofern sie schwache Verben sind, auf folgende Endungen geprüft werden:

Nomenklatur: P = Person s = Singular p = Plural i = Indikativ k = Konjunktiv

Präsens (Indikativ + Konjunktiv)

- e	(1Psi, 1PskI, 3PskI)
-[e]st	(2Psi)
- e st	(2PskI)
- e n	(1Ppi, 3Ppi, 1PpkI, 3PpkI)
-[e] t	(3Psi, 2Ppi)
- e t	(2PpkI)

Präteritum (Indikativ + Konjunktiv II)

- [e] t e	(1PsikII, 3PsikII)
- [e] t e n	(1PpikII, 3PpikII)
- [e] t e s t	(2PsikII)
- [e] t e t	(2PpikII)

Sortiert man diese Marker, so ergeben sich folgende Regeln zur Findung des Verbstamms:

1. [e][t][e]	{ete, te, e}
2. [e][t][en]	{eten, ten, en}[6]

3. [e][te, e][t] {etet, tet, et, t}
4. [e][te, e][st] {etest, test, est, st}

b) starke Verben

Wie die schwachen Verben so bilden auch die starken Verben ihre finiten Formen mit den Flexionssuffixen -e, -st, -t und -en. Unterschiede bestehen nur bei der Bildung des Präteritums und des Partizip Perfekts:

- die 1. und 3. Person Sing. Indikativ Präteritum sind endungslos
- die Endungen der anderen Personalformen entsprechen denen des Präsens

Die starken Verben bilden den Infinitiv mit der Endung -en, das Partizip I mit der Endung -end und das Partizip II mit der Vorsilbe ge- und der Endung -en.

Somit müssen folgende Endungen geprüft werden:

Präsens (Indikativ und Konjunktiv)

(entspricht den bei den schwachen Verben genannten Formen; siehe oben)

Präteritum

-	(1Psi, 3Psi)
- e	(1PskII, 3PskII)
- [e] s t	(2PsikII)
- e n	(1PpikII, 3PpikII)
- [e] t	(2PpikII)

Sortiert man diese Marker, so ergeben sich für das Präteritum der starken Verben folgende Regeln:

5. [ ] {leere Endung}
6. [e] {e}
7. [en] {en}
8. [e][t] {et, t}
9. [e][st] {est, st}

Da die Regel 5 (leere Endung) schon beim ersten Aufsuchen des Lexikons gefunden wird und die Regeln 6-9 in dem Regelsatz der schwachen Verben enthalten sind, genü-

gen also Regel 1-4, um alle finiten Verbformen auf ihren Verbstamm zurückzuführen. Im Lexikon sind die drei Stammformen für Präsens, Präteritum und Partizip II sowie die Ablautform für 2. und 3. Person Singular Präsens aufgenommen.
Beispiel:

sprechen:	sprech - sprich - sprach - sproch
kommen:	komm - kam
sehen:	seh - sieh - sah
nehmen:	nehm - nimm - nahm - nomm

Umlaut

Sofern der Verbstamm einen umlautfähigen Vokal enthält, tritt Umlaut an zwei Stellen auf: in der 2. und 3. Person Singular Indikativ Präsens und in allen Formen des Konjunktivs II. Seltene Formen wie "ich begönne" oder "ich hübe" (Konj. II von heben) sind nicht im Lexikon aufgenommen.

Bei der Verbanalyse werden alle übergebenen Lemmata auf Umlaut untersucht und entsprechend geflaggt. Wenn also trotz Reduzierung der Stamm im Lexikon nicht gefunden wird und die Umlaut-Flagge gesetzt ist, wird der Umlaut zurückgesetzt und diese Form im Lexikon gesucht. Bleibt auch das erfolglos, wird das Lemma als "unbekannt" markiert und die Verbanalyse verlassen.

Es gibt aber einige Fälle, wo ohne syntaktische Analyse nicht eindeutig die Stammform bestimmt werden kann:

(13) An seiner Stelle führe ich nicht nach Berlin. (Konj. II von "fahren")
(14) Heute Nachmittag führe ich dich durch die Altstadt. (Indikativ von "führen")

Um solche Konflikte zu lösen, müssten den Verben im Lexikon Angaben zu ihrer Valenz hinzugefügt sowie die Lemmata eines Satzes auf ihre syntaktische Funktion untersucht werden. Obwohl mir solch eine Ausweitung des Programms prinzipiell wünschenswert erscheint, wurde sie aus Zeitgründen zurückgestellt.

c) gemischt konjugierte Verben

Für diese Verbgruppe wird eine regelmäßige Konjugation angenommen. Daher ist ihre Analyse durch die Regeln der schwachen Konjugation gedeckt. Wie bei den star-

ken Verben sind die entsprechenden Stammformen ins Lexikon aufgenommen:

denken:	denk - dach
bringen:	bring - brach
kennen:	kenn - kann

Eine Lösung für Konflikte wie "kann" (können) und "kann-te" (kennen) steht noch aus.

d) Behandlung der Partizipien

Für die Analyse der Partizipien wurde eine eigene Unterroutine entwickelt, die von der Verbanalyse, aber auch von der Adjektivanalyse aufgerufen wird, da ja Partizipien auch attributiv verwendet werden. Die Merkmale der Partizipien sind:

- Suffix -end für das Partizip Präsens (alle Verbgruppen)
- Präfix bzw. Infix -ge- für das Partizip Perfekt (alle Verbgruppen; Ausnahmen s.u.)
- Suffix -en (starke Verben) oder -t (schwache bzw. gemischt konj. Verben)

Kein -ge- haben Verben mit nicht trennbarem Präfix und Verben auf -ieren:

vergehen - vergangen	(Lexikon: vergeh - verging - vergang)
verkaufen - verkauft	(Lexikon: verkauf)
studieren - studiert	(Lexikon: studier)
einstudieren - einstudiert	(Lexikon: studier, einstudier)

Diese Formen werden (teilweise syntaktisch falsch) als finite Verbform erkannt. Da die Zuordnung zum Verb aber richtig ist, wurde der Zustand zunächst beibehalten.

Zusammenfassend ergeben sich für die Analyse der Partizipien folgende Regeln:

- [end]	Partizip I
- [ge]...[t, en]	Partizip II

6. Programmdurchlauf

Nachdem nun alle Module des Programms besprochen sind, soll als letztes an einem

kleinen Text die Ausgabe des Programms demonstriert werden.

Zunächst der Test-Text[7]:

Das Geld oder das Leben

Herr Petermann ging allein im Wald spazieren. Da sprang ein Mann auf ihn zu, hielt ihm einen Revolver vor die Nase und schrie: "Das Geld oder das Leben!"
Herr Petermann erschrak natürlich sehr. Er hatte keinen Revolver und konnte auch nicht schießen, aber er verlor seine Ruhe nicht und sagte freundlich zu dem Räuber: "Mein Herr, ich gebe Ihnen lieber mein Geld als mein Leben. Aber ich fürchte mich, ohne Geld nach Hause zu kommen. Was soll ich meiner Frau sagen? Bitte, helfen Sie mir! Schießen Sie mir ein Loch durch die Jacke; dann muß meine Frau glauben, was ich erzähle."

[... Text abgekürzt]

Zum Lex-Level 5 wird folgendes ausgegeben:

1: Petermann*	10: Räuber*	19: Kugel*
2: spazieren*	11: fürchte*	20: ärgerte*
3: sprang	12: Schießen*	21: Spaziergang
4: Revolver*	13: Jacke	22: Petermanns*
5: schrie	14: solch	23: klug
6: erschrak*	15: schoß*	24: Spiel
7: schießen*	16: vorsichtig*	25: gewonnen
8: verlor	17: tat	
9: Ruhe	18: Schuß*	

* : nicht im Lexikon enthaltene Wörter

Text-Info

31 Sätze 6 Abschnitte 277 Wörter

25 gelistete Wörter (9.02%) 10 Wörter höheren Levels (3.61%)

15 unbekannte, im Lexikon nicht enthaltene Wörter (5.41%)

Lex-Level: 5 (1670 Wörter befinden sich im Lexikon)

Interpretiert man nun die Daten dieser Ausgabe, so ergibt sich folgendes Bild:

- der Schwierigkeitsgrad 5 setzt schon 800 Wörter als bekannt voraus!
- der Text enthält unbekannte, im Lexikon nicht aufgenommene Wörter
- Familiennamen und andere Eigennamen sind zur Zeit ebenfalls nicht im Lexikon aufgenommen, können aber aus dem Textkontext erschlossen werden
- markierte Wörter wie "Revolver", "Räuber", "Kugel" sind schwer wegzulassen oder durch andere Wörter zu ersetzen.

Das zeigt, wie schwierig es ist, einen Text für Leseanfänger zu schreiben, der interessant ist, aber trotzdem den eingeschränkten Wortschatz dieser Lerner berücksichtigt!

C Ausblick

Der Leser wird vielleicht überrascht sein, dass selbst auf der Stufe 5 noch 25 Wörter aufgelistet werden und die für leichte Lesbarkeit anvisierte 5%-Hürde um weitere 4% überschritten wird. Ein "native speaker" der deutschen Sprache wird vielleicht gedacht haben: "Aber der Text ist doch ganz einfach!" - Ja, für Muttersprachler sicherlich, nicht jedoch für Deutsch als Fremdsprache lernende Menschen anderer Länder.
Ebenso ist erstaunlich, dass das Verb "erschrecken / erschrak" in der Top-10.000-Wortliste der Universität Leipzig nicht enthalten ist und das Verb "schießen" erst an 5828. Stelle rangiert! Dagegen findet sich das ein wenig "altmodisch" anmutende Wort "Räuber" auf Platz 6712.

Durch dieses Programm kann also eine Lehrperson eher beurteilen, wie schwer denn nun ein Text wirklich für die Lerner ist, denn das eigene Gefühl täuscht nur allzu oft.
Das Programm gestattet dem Lehrer außerdem, die Lex-Stufe zu finden, auf der der Text "leicht" und ohne Überforderung gelesen werden kann. Auf diese Weise können Lerner einen ihrer Wissensstufe entsprechenden Text auswählen und vielleicht dabei erfahren, dass Lesen "Spass" machen kann.
Jedoch sind gerade für Anfänger geeignete Texte mit einer Wortschatzanforderung von 300-400 Wörtern kaum erhältlich. Hier bietet es sich an, einmal selbst leichte Texte zu verfassen. Wieder leistet das LEKTOR-Programm bedeutende Hilfestellung: einerseits kann man feststellen, welche Wortschatzanforderungen der Text an die Lerner stellt, andererseits kann

man durch die Ausgabe der Wörter, die den anvisierten Level übersteigen, diese gezielt vermeiden bzw. durch andere, einfachere ersetzen.

Das Programm ist noch im Alpha-Stadium, es befindet sich in der Entwicklung und ist an vielen Stellen zu verbessern. Auch das Lexikon befriedigt in seiner jetzigen Form nicht ganz. Folgende Punkte stehen dabei zur Bearbeitung an:

- nochmalige Überarbeitung der Rangfolge der Wörter
- Aufstockung des Lexikons auf ca. 2000 Wörter
- feinere Differenzierung der Wortklassen
- mehr grammatisch-syntaktische Merkmale pro Lemma

Das Programm wurde trotz seiner Unvollkommenheit schon im Unterricht eingesetzt und hat seine Nützlichkeit bewiesen. Der Verfasser hofft, es nach der Überarbeitung anderen, an dieser Thematik interessierten Kollegen zur Verfügung stellen zu können.

verwendete und zitierte Literatur

- DUDEN Grammatik, 5. Auflage, Mannheim 1995
- Neuner/Hunfeld, Methoden des fremdprachlichen Deutschunterrichts, Kassel 1993
- Swantje Ehlers, Lesen als Verstehen, Kassel 1992
- Elke Donalies, Die Wortbildung des Deutschen, Tübingen 2005
- Hans Eggers, Elektronische Syntaxanalyse des Deutschen, Tübingen 1969
- Oliver Lorenz, Automatische Wortformerkennung für das Deutsche im Rahmen von MALAGA (Magisterarbeit), Universität Erlangen-Nürnberg 1996
- Die Wortformen der geschlossenen Wortarten im Stuttgart-Tübingen Tagset (STTS), Universität Tübingen 1996
- Wolfgang Lezius, Die Wortklassensysteme von Morphy, interner Bericht, Universität Paderborn 1998
- Clematide, S./Volk, M., Linguistische und semantische Annotation eines Zeitungskorpus, GLDV-Jahrestagung, Giessen 2001
- Gamon/Reutter, The Analysis of German Separable Prefix Verbs in the Microsoft Natural Language Processing System, Microsoft Research, Redmont 1997
- Das GTU-Lexikon, Universität Koblenz o. Jahr
- Martin Grund, GUT1-Handbuch, Computer & Lernen, Baden-Baden 2004

- Robert Waring, At what rate do learners learn and retain new vocabulary from reading a graded reader?, Reading in a Foreign Language, Volume 15, No. 2, Hawaii 2003
- Hermann Kessler, Deutsch für Ausländer, Leichte Erzählungen, Königswinter 1973

Anmerkungen

1 Waring, S. 153/154

2 Diese am Institut für Sprachverarbeitung der Universität Leipzig erstellten Listen sind zwar noch unter folgender URL downloadbar: http://wortschatz.uni-leipzig.de/html/wliste.html, eine Referenz dazu findet sich aber weder auf der Seite des Instituts noch auf der des Wortschatz-Portals.

3 GUT1-Handbuch, S. 34-35

4 vergl. dazu den Aufsatz von Gamon/Reutter

5 Beispiele laut Duden, S. 233-234

6 -en wird auch für den Infinitiv berücksichtigt

7 Dieser Text ist dem Heftchen "Deutsch für Ausländer", S. 16, entnommen.

（Ruediger Schmidt　国際言語学部教授）