

Datenbankbasierte Verwaltung und Pflege morphologischer Information im IMSLex

Wolfgang LEZIUS, Arne FITSCHEN, Ulrich HEID

1. Einführung

Am Institut für Maschinelle Sprachverarbeitung sind in den vergangenen Jahren eine Reihe lexikalischer Ressourcen für das NLP entstanden. Dazu zählen u.a. das Lexikon einer Morphologie-Komponente (SCHILLER 1996), ein Tagger-Lexikon (SCHMID 1994), ein Subkategorisierungslexikon (ECKLE 1998, ECKLE/HEID 1996) und weitere für das Syntax-Parsing verwendete Lexika (BRÖKER/DIPPER 1999). Um eine wechselseitige Verzahnung in einer einheitlich zugänglichen Quelle zu erreichen, wurde das IMSLex entwickelt. Es handelt sich hierbei um eine relationale Datenbank, die die einzelnen Teillexika über Datenbank-Tabellen miteinander verbindet. In diesem Papier wird der Teilbereich Morphologie erläutert, der Voraussetzung für den Aufbau der übrigen Module (z.B. des Syntax-Lexikons) ist.

Die Verwaltung der lexikalischen Daten steht damit im Gegensatz zu den meisten verfügbaren Morphologie-Systemen des Deutschen, die ihre Daten in dateibasierten Datenformaten ablegen (vgl. z.B. das Morphy-System, LEZIUS 1998). Letztere Vorgehensweise erleichtert die Anbindung und steigert die Effizienz. Doch die Verwendung einer Datenbank bringt zusätzlich die Garantie der Aktualität der Daten, da die Anwendungen stets auf den aktuellen, zentral verwalteten Datenbestand zugreifen (vgl. auch das TransLexis-System, BLÄSER 1998). Durch die Verfügbarkeit von Datenbank-Schnittstellen für alle gängigen Programmiersprachen kann die Datenbank zudem problemlos in weitere Anwendungen eingebunden werden.

2. Pflege des Morphologischen Datenbestands

Beim Aufbau des Morphologischen Datenbestands lag es nahe, zunächst den vorhandenen Datenbestand der Morphologie-Komponente in

die Datenbank zu überführen. Daneben wurde die phänomenbasierte Erweiterung des Bestands auf der Grundlage von Textkorpora vorangetrieben. Für die Pflege einzelner Stämme schließlich wurde versucht, ein benutzerfreundliches Konzept zur interaktiven Lexikonpflege zu entwickeln. Diese Methoden zum Aufbau und zur Pflege des Datenbestands werden im folgenden erläutert.

2.1 Migration 2-Ebenen-Morphologie zur Datenbank

2.1.1 Die Ressource: 2-Ebenen-Morphologie

DMOR, die deutsche Morphologie, ist eine Sammlung von Daten, die sich zu einem schnellen Analyse- und Generierungswerkzeug für die deutsche Sprache kompilieren lassen. Sie basiert auf der Idee der 2-Ebenen-Morphologie von Koskeniemi (vgl. SCHILLER 1996).

Die Quell-Dateien zur Erzeugung des Werkzeugs (endlicher Automat) teilen sich auf in Stammlisten mit zugehörigen Flexionsklasseninformationen auf der einen, Fortsetzungsklassen und Regelninformationen auf der anderen Seite. Erstere Dateien („Stammdateien“) dienen dazu, Wortstämme und ihre Flexionsklassen über Fortsetzungsklassen in Flexionsparadigmen zu expandieren. Die Fortsetzungsklassen geben zu einem Wortstamm alle Flexionsformen mitsamt der morphologischen Informationen an. Die 2-Ebenen-Regeln haben die Aufgabe, aus den erzeugten Vollformen die unerwünschten auszuschließen bzw. in der Vollform bestimmte Zeichen hinzuzufügen, zu ändern oder zu entfernen (z.B. bei der e-Elision).

Der Aufbau einer jeden Stammdatei folgt einem festgelegten Muster. Zwei Einheiten definieren im wesentlichen¹ einen Eintrag, der Stamm und die zugehörige Flexionsklasse.

Beispiel: Aa1 NMasc_es_e;

Diese Informationen allein ermöglichen zusammen mit den 2-Ebenen-Regeln die Generierung der Vollformen zu jedem gespeicherten Stamm.

¹ Hinzu kommt noch die sogenannte „Oberflächenform“ für Suppletivformen.

2.1.2 Das Resultat: Die Datenbank

Für die morphologische Datenbank wird auf die Speicherung der 2-Ebenen-Regeln verzichtet, da die Regelanwendung in der Datenbank nicht operationalisierbar ist. Stattdessen wird der wesentliche Informationsgehalt der Stammdateien festgehalten, um aus der Datenbank und den 2-Ebenen-Regeln, die separat gehalten werden, den endlichen Automaten zur Analyse und Generierung erzeugen zu können. Dies erscheint sinnvoll, da die Stämme laufend erweitert werden, die Regeln (Pluralumlautung, e-Elision etc.) jedoch weitgehend statisch sind.

Neben den notwendigen Feldern kommen einige zusätzliche Angaben hinzu, die die Verwaltung der Datenbank unterstützen bzw. weitere Anwendungen ermöglichen, beispielsweise die Angabe der Wortart für die Ausgabe aller gespeicherten Formen in ein Taggerlexikon. Als Grundtabelle dient die abstrakte Tabelle `IMS_Lexikon` (vgl. Abbildung 1).

Feldname	Typ	Länge	Bemerkung
Lemma	Text	variabel	
Oberflächenform	Text	variabel	Suppletivform: ging → geh(en)
Flexionsklasse	Text	variabel	nach DMOR-Spezifikation
Part_of_Speech	Text	variabel	nach STTS-Tagset
Name_Eintragender	Text	variabel	als Kürzel
Name_Administrator	Text	variabel	als Kürzel
Eintragsdatum	Datum	fix	z.B. 03/23/1999
Eintragsvermerk	Text	variabel	Phänomen-Markierung
Kommentar	Text	variabel	für spätere Kommentare

Abbildung 1: Die Grundtabelle `IMS_Lexikon`

In den weiteren Tabellen wird zusätzlich nach Wortart unterschieden. Diese Einteilung geht ursprünglich auf die verschiedenen Stammdateien zurück, ist aber auch sinnvoll, wenn es darum geht, Daten verschiedener Wortarten miteinander zu vergleichen. Beispiel: Läßt man sich alle Verbstämme ausgeben und mit der Endung *-bar* versehen, dann kann man

direkt in der Datenbank durch einen Vergleich mit der Adjektivtabelle ersehen, welche Adjektive auf *-bar* von Verben abgeleitet sein könnten. Weiterhin ist es erst so möglich, wortartenspezifische Informationen, z.B. zur Partizipbildung bei Verben oder zur Verwendung von Adjektiven, nur dort unterzubringen, wo sie tatsächlich anfallen.

Für die weitere Entwicklung der IMSLex-Datenbank ist vorgesehen, parallel zum morphologischen Teil der Datenbank einen semantischen und einen syntaktischen Teil aufzubauen, die über eine Lemma-Tabelle miteinander verbunden werden.

2.2 Korpusbasierte Erweiterung des Datenbestands

Anhand von Zeitungskorpora wurde der durch die Migration der 2-Ebenen-Morphologie verfügbare Datenbestand systematisch um Material zu Derivation und Komposition erweitert.

Dabei besteht das Ziel darin, mittelfristig den Anschluß einer Komponente zur Derivations- und Kompositionsmorphologie zu ermöglichen, die regelhaft analysierbare komplexe Wörter automatisch zerlegt, ohne sie im Lexikon halten zu müssen. Kurzfristig müssen jedoch auch solche Wortbildungsprodukte im IMSLex erfaßt und zusammen mit einer Phänotyp-Markierung abgelegt werden. Diese Markierung erlaubt den späteren modulweisen Ersatz der Lemma-Gruppen durch die Ergebnisse einer Wortbildungskomponente.

Die benutzten Korpora sind "opportunistisch" gesammelt worden (vgl. die Tabelle in Abbildung 2); für die Erweiterung des Lemmabestandes des IMSLex führt dies aber nicht zu Problemen.

Korpus	Zeitraum	Jgge.	Umfang	(Wf.)
Frankfurter Rundschau	1992/93	2	40	Mio.
Stuttgarter Zeitung	1992/93	2	36	Mio.
die tageszeitung	1987-93	7	103	Mio.
'European News Cp.': dpa, afp...	1990-94	5	100	Mio.

Abbildung 2: Benutzte journalistische Korpora

Das verwendete Verfahren macht sich den Umstand zunutze, daß Wortstämme gleicher Endung in der Regel der gleichen Flexionsklasse angehören. Mithilfe des Korpus-Anfrage-Systems CQP (CHRIST 1995) wurden Listen von Wortformen- bzw. Lemmakandidaten extrahiert, die dieselben Endungen aufweisen, z.B. Adjektivkandidaten auf *-bar*, *-lich*, *-ig* usw.

Durch Abgleich gegen den Lemmabestand des IMSLex wurde die Liste der bisher in IMSLex nicht bekannten Kandidaten ermittelt. Diese wurden manuell bzw. semi-automatisch hinsichtlich Phänomentyp und Flexionseigenschaften klassifiziert und gruppenweise in das IMSLex aufgenommen. Auf diese Weise konnte der Lemmabestand in relativ kurzer Zeit (ca. 2 Personen-Monate) um ca. 15.000 neue Lemmata ergänzt werden.

Die phänomenologische Klassifikation ist hier insofern besonders relevant, als ein erheblicher Anteil der aus Zeitungstext extrahierten Kandidaten aus transparenten Komposita (die nicht im IMSLex angegeben werden), Ad-hoc-Komposita, orthographischen Fehlern oder Eigennamen besteht. Daneben treten Fälle auf, für die klare deskriptive Richtlinien erforderlich sind. Hierzu gehören lexikalisierte Formen, bei denen die aus der Oberflächenform abgeleitete Hypothese bezüglich eines derivationellen Bildungsmusters nicht zutrifft: *Kittchen*, *Frettchen*, *Mätzchen*, *Flittchen* sind gerade nicht als Verkleinerungsformen zu klassifizieren.

2.3 Ein Konzept zur interaktiven Lexikonpflege

Für den weiteren kontinuierlichen Ausbau des Lexikons ist eine eigene, interaktive Komponente entwickelt worden. Um dessen Konzept besser einordnen zu können, wird im folgenden ein Überblick über Ansätze gegeben, die sich in der Praxis bewährt haben.

2.3.1 Bisherige Ansätze

Die meisten Morphologie-Systeme sehen keine Unterstützung des Anwenders bei der Lexikon-Akquisition vor. Die Pflege ist meist nur durch das Editieren kryptischer Quelltexte möglich. Dieser Umstand macht

ein System, mag es die Sprachphänomene noch so gut abdecken, für andere Anwender nicht einsetzbar.

Ein alternativer Ansatz wurde u.a. im Morphy-System verfolgt (vgl. LEZIUS 1998): Der Anwender wird nach Eingabe des Wortstamms durch einen Dialog geführt, in dessen Verlauf er Fragen bezüglich des Flexionsparadigmas beantworten muß. Dabei werden stets Wortformen als Alternativen vorgeschlagen, die richtige muß lediglich ausgewählt werden. Die Implementation des Dialogs stützt sich auf einen Entscheidungsbaum, an dessen Knoten die Fragen nach bestimmten Phänomenen stehen, die für eines der möglichen Flexionsparadigmen charakteristisch sind. Die Kanten sind mit den Antwortmöglichkeiten einer solchen Frage beschriftet und werden abhängig von der Auswahl des Benutzers beschriftet. Führt der Dialog zu einem Blatt und damit zum Ziel, so ist die dort abgelegte Flexionsklasse die gesuchte. Ein Beispiel für einen Entscheidungsbaum und einen entsprechenden Benutzer-Dialog findet sich bei der Vorstellung des IMS-Ansatzes (vgl. 2.3.2).

Ein weiterer Ansatz verfügt über Wissen darüber, welche Flexionsklassen für einen Stamm mit einer spezifischen Endung möglich sind. So ist beispielsweise die Flexion eines Nomens mit der Endung *-keit* stets eindeutig. Für jeden neuen Stamm werden so alle möglichen Flexionsklassen generiert und nach ihrer Auftretenshäufigkeit sortiert vorgeschlagen. Gibt es keine passende Alternative, gibt der Benutzer das Paradigma Wortform für Wortform ein. Da sich die Flexionsklassen für spezifische Suffixe speziell im Deutschen sehr ungleichmäßig verteilen (einige Klassen sind sehr häufig, die übrigen eher selten), ist dieses Verfahren sehr effektiv und wurde bereits erfolgreich bei der Pflege des CISLEX-Lexikons eingesetzt (vgl. MAIER 1998).

2.3.2 Der Ansatz des IMS

Zielsetzung beim Entwurf eines Pflegekonzepts war die Erweiterung des Lexikons durch Nicht-Fachleute. Basis unseres Ansatzes ist ein Entscheidungsbaum, der um Funktionen zur Bestimmung der wahrscheinlichsten Klasse erweitert wurde: Liegt ein Suffix vor, das fast ausschließlich mit einer bestimmten Flexionsklasse einhergeht, wird vorab eine spezifi-

sche Frage gestellt, die einen eindeutigen Rückschluß auf diese Klasse zuläßt. Die Integration dieses Ansatzes befindet sich jedoch noch im Anfangsstadium.

Für die Pflege wurden Entscheidungsbäume für Substantive, Eigennamen, Adjektive und regelmäßige Verben entworfen. Für die übrigen Wortklassen ist die manuelle Angabe der Flexionsklasse erforderlich. Die Abbildungen 3 und 4 zeigen den Entscheidungsbaum für Adjektive sowie den Baumdurchlauf für das Adjektiv alt (Benutzereingaben sind unterstrichen).

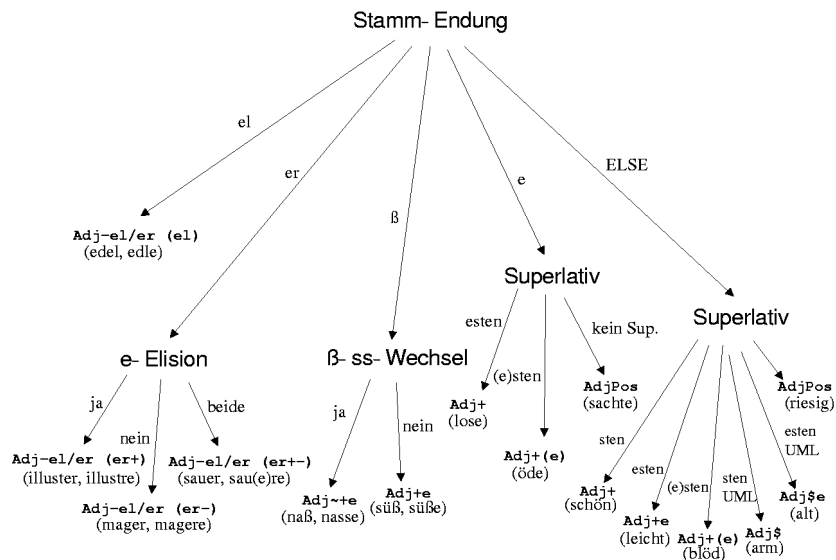


Abbildung 3: Entscheidungsbaum für Adjektive

2.3.3 Die Architektur der Akquisitionskomponente

Neben dem Pflegekonzept spielt das Konzept der Umsetzung für die Praxistauglichkeit eine große Rolle. Der folgende Abschnitt skizziert daher die Systemarchitektur.

<ol style="list-style-type: none">1. Geben Sie den Stamm ein: <i>alt</i>2. Welche Wortklasse liegt vor?<ol style="list-style-type: none">1: Substantiv2: Eigename3: Adjektiv4: Verb3. Wie lautet die Superlativ-Form?<ol style="list-style-type: none">1: am altsten2: am altesten3: am alt(e)sten4: am ältsten5: am ältesten6: kein Superlativ <p>Klasse Adj\$e ermittelt!</p>

Abbildung 4: Benutzerdialog beim Eintrag des Adjektivs *alt*

Für die Implementation ist der Entscheidungsbaum-Dialog eingebettet in eine Client/Server-Architektur (vgl. Abbildung 5). Ein Serverprozeß steuert die Anbindung der Datenbank und eines Moduls zur Generierung von Flexionsparadigmen. Diese dienen der Kontrolle neuer Benutzereinträge. Die Clients werden dadurch von plattformspezifischen Bestandteilen befreit. Durch die Verwendung von Java als Entwicklungssprache sind die Clients plattformübergreifend und durch die Verwendung des TCP/IP-Protokolls sogar über Rechnernetze hinweg einsetzbar.

Der Eintrag eines neuen Stamms vollzieht sich schrittweise (vgl. Numerierung in Abbildung 5). Zunächst übergibt der Client den Stamm an den Server, um zu prüfen, ob der Stamm bereits in der Datenbank eingetragen ist (1). Ist dies nicht der Fall bzw. liegt eine weitere Lesart vor, läuft der interaktive Dialog an, an dessen Ende die Flexionsklasse ermittelt ist (2). Diese Aufgabe wird allein vom Client geleistet. Sie ist wenig rechenintensiv und beschränkt die Client-Server-Kommunikation auf das Wesentliche. Zur Kontrolle der ermittelten Klasse wird das Flexionsparadigma des Stamms generiert (3). Diese Funktion leistet das serverseitige Perl-Modul *AMOR*.

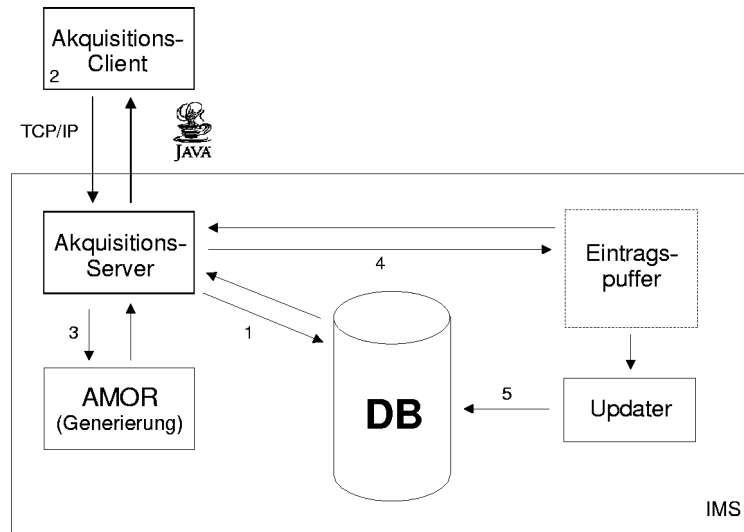


Abbildung 5: Client-/Serverarchitektur der Akquisitionskomponente

Stimmt das gezeigte Paradigma mit den Vorstellungen des Benutzers überein, wird der Stamm im sogenannten Eintragspuffer zwischengespeichert (4) — er gelangt also nicht direkt in die Datenbank. Die Übertragung in die Datenbank erfolgt in einem separaten Schritt durch ein Update-Tool, das durch die Datenbank-Administratoren kontrolliert wird (5). Diese aktualisieren in regelmäßigen Abständen die Datenbank. Vorteile einer Zwischenspeicherung sind die Kontrollmöglichkeit der Einträge für Administrator und Benutzer vor der nächsten Aktualisierung.

3. Zusammenfassung und Ausblick

Wir haben ein Konzept und die Realisierung einer morphologischen Datenbank vorgestellt. Der Datenbestand ist mit Hilfe der drei Zugänge Migration aus einer bestehenden Morphologie-Komponente, korpusbasierte Pflege und interaktive Lexikonpflege aufgebaut worden. Das morphologische Lexikon umfaßt derzeit 65.000 Lemmata.

Für die Teildatenbank Morphologie steht neben der weiteren kontinuierlichen Pflege in der Zukunft die Weiterentwicklung der Regelkomponente im Vordergrund. Durch die phänomenbasierte Lexikonerweiterung

steht aussagekräftiges Anschauungsmaterial für die Behandlung der Komposition und Derivation zur Verfügung.

Für das IMSLex soll vor allem die Integration der Teildatenbanken Morphologie und Syntax vorangetrieben werden.

Literatur

- BLÄSER, B. (1998): TransLexis: An Integrated Environment for Lexicon and Terminology Management. IBM Heidelberg, Arbeitsbericht.
- BRÖKER, N. / DIPPER, S. (1999): Zur Konstruktion von Lexika für die maschinelle syntaktische Analyse. (In diesem Band).
- CHRIST, O. (1994): A modular and flexible architecture for an integrated corpus query system. In: *Proceedings of the COMPLEX 1994*, 23-32.
- ECKLE, J. (1998): Methods for quality assurance in semi-automatic lexicon acquisition from corpora. In: *Proceedings of EURALEX 1998*, 119-128.
- ECKLE, J. / HEID, U. (1996): Extracting raw material for a German subcategorization lexicon from newspaper text. In: *Proceedings of the 4th International Conference on Computational Lexicography, COMPLEX 1996*, 39-51.
- LEZIUS, W. (1999): Das IMSLex - Online-Informationen. <http://www.ims.uni-stuttgart.de/projekte/IMSLex/>
- LEZIUS, W. / RAPP, R. / WETTLER, M. (1998): A Freely Available Morphological Analyzer, Disambiguator, and Context Sensitive Lemmatizer for German. In: *Proceedings of the COLING-ACL 1998*, 743-748.
- MAIER, P. (1998): Defaultzuweisung morphosyntaktischer Kategorien. In: HEYER, G. / WOLFF, C. (eds.): *Linguistik und Neue Medien*, 151-162. Wiesbaden: Deutscher Universitätsverlag.
- SCHILLER, A. (1996): Deutsche Flexions- und Kompositionsmorphologie mit PC-KIMMO. In: HAUSSER, R. (ed.): *Linguistische Verifikation – Dokumentation zur Ersten Morpholympics 1994*, 37-52. Tübingen: Niemeyer.
- SCHMID, H. (1994): Probabilistic Part-of-Speech Tagging Using Decision Trees. In: *Proceedings of the International Conference on New Methods in Language Processing 1994*, 44-49.