Erschienen in: Lobin, Henning (Hrsg.): Sprach- und Texttechnologie in digitalen Medien. Proceedings der GLDV-Frühjahrstagung 2001. 28.-30. März 2001 Justus-Liebig-Universität Gießen. – Norderstedt: Libri Books on Demand, 2001. S. 41-48.

Lexikonexpansion – Vom XML-annotierten Stammformenlexikon zum Vollformenlexikon*

Jens Pönninghaus und Andreas Witt

Zusammenfassung

Im folgenden wird eine texttechnologische Komponente zur Expansion eines XML-annotierten Stammformenlexikons, das auf Einträgen eines Standardwörterbuchs basiert, vorgestellt. Diese Expansion wurde in der Document Style Semantics and Specification Language implementiert. Ihr Ergebnis ist ein Vollformenlexikon, das ebenfalls in XML repräsentiert ist.

4.1. Einleitung

Lexika stellen für die Linguistik eine wichtige Informationsquelle dar. Im Gegensatz zu den meisten sprachtechnologischen Anwendungen, die auf spezialisierte computerlinguistische Lexika angewiesen sind, wird in der nachfolgend beschriebenen Implementierung ein klassisches, maschinenlesbares Wörterbuch als Lexikon verwendet.[1]

Der Vorteil der Nutzung klassischer Wörterbücher gegenüber der Verwendung spezialisierter computerlinguistischer Lexika für die Texttechnologie besteht insbesondere darin, dass sie eine Ressource bilden, die für eine Vielzahl von Sprachen zur Verfügung steht und dass einige dieser Lexika den Anspruch besitzen, das Lexeminventar einer Sprache zu einer bestimmten Zeit möglichst vollständig aufzuführen. Dadurch werden hochvolumige Wissensquellen erschlossen, deren Inhalte durch jahrelange Pflege gut validiert sind.

Die Aufgabe der hier vorgestellten Komponente zur Lexikonexpansion ist die Überführung eines in XML-Notation vorliegenden Stammformenlexikons in ein Vollformenlexikon, das ebenfalls in XML notiert sein soll. Die Abbildung wird mit Hilfe einer Dokumentsemantik definiert. Es ergeben sich grundsätzlich zwei Möglichkeiten diese Transformation zu realisieren: Entweder dynamisch während der Zugriffszeit auf die Vollformen oder aber statisch in einem Offline-Verfahren. Die Entscheidung zugunsten der Offline-Variante erfolgt aus Gründen der zeiteffizienten maschinellen Verarbeitung relevanter Daten mit texttechnologischen Methoden.

Für andere Anwendungsgebiete ist die Methode einer dynamischen Expansion, wie sie beispielsweise in den Beiträgen zur Morpholympics (vgl. Hausser, 1996) vorgestellt werden, zweifelsohne vorzuziehen.

* Erschienen in: *Proceedings der GLDV-Frühjahrstagung 2001*, Henning Lobin (Hrsg.), Universität Gießen, 28.–30. März 2001, Seite 41–48. `http://www.uni-giessen.de/fb09/ascl/gldv2001/`

[1] Die vorgestellte Komponente stellt einen Teil einer Diplomarbeit zur automatischen Annotierung SGML/XML-strukturierter Dokumente dar (Pönninghaus, 2000).

Eine Einschränkung des nachfolgend vorgestellten Ansatzes besteht darin, dass er nur für überwiegend flektierende Sprachen sinnvoll erscheint. Für agglutinierende Sprachen mit ihrer ungünstigen Ratio von Vollformen zu Stammformen ist dieser Ansatz vermutlich nur in restringiertem Maße sinnvoll, z. B. bei einem eingeschränkten Lexikon.

4.2. Systemarchitektur

Wie aus Abb. 4.1 ersichtlich, erfolgt die Generierung der Vollformen mit Hilfe der Document Style Semantics and Specification Language (DSSSL, ISO/IEC, 1996). Sie ist eine Sprache zur Beschreibung von Semantiken für SGML-Dokumente, die auf der Programmiersprache Scheme basiert. DSSSL ist konzeptionell mit der neueren Extensible Style Language (XSL) vergleichbar, d. h. sie besteht aus einer Transformations- und einer Formatierungssprache und integriert eine seiteneffektfreie, funktionale Programmiersprache. Da DSSSL im Gegensatz zu XSL auf einer in der Computerlinguistik etablierten Programmiersprache aus der Lisp-Familie basiert, ist es vorteilhaft, sie für die Verarbeitung SGML- oder XML-repräsentierter natürlichsprachlicher Daten zu verwenden (vgl. Witt, 1999).

Die Generierungskomponente bildet die Lexeme eines XML-strukturierten Eingabelexikons gemäß eines definierbaren Paradigmensatzes auf die gewünschten Vollformen ab. Die Größe des Bildbereiches der Abbildung, d. h. welche Klassen der Vollformen generiert werden, wird durch den Paradigmensatz gesteuert.

Gegenstand der Abbildung sind nur die Prozesse der Flexion. Dies hat zur Folge, dass die im Deutschen recht häufig vorkommenden Wortbildungen – also Derivation und Komposition – explizit ausgeschlossen werden, da diese Prozesse ihrerseits zu Stammformen führen, die regelmäßig bzw. unregelmäßig flektiert werden, also Teil der Eingabe der Generierungskomponente sind. Viele gebräuchliche Komposita und Derivative sind zudem ohnehin bereits lexikalisiert, d. h. in einem Standardwörterbuch als eigenständige Einträge aufgeführt. Insbesondere für das Deutsche mit seinen vielfältigen Möglichkeiten der Wortbildung durch Komposition und Derivation ist es allerdings sinnvoll, entsprechende Komponenten zu implementieren und dem Vollformengenerierungsprozess vorzuschalten. Dies ist derzeit nicht implementiert, jedoch durch die modulare Gestaltungsweise des Systems jederzeit integrierbar.

Die Regeln der deutschen Morphotaktik sind innerhalb des Generierungsalgorithmus nachgebildet. Die Flexive (z. B. „t" und „e" in „back+t+e") und die unregelmäßigen Formen („backen" „buk") werden in separaten Datenstrukturen vorgehalten.

Abb. 4.2 zeigt den inneren Aufbau der Generierungskomponente. Sie besteht aus der Steuerungs-, der Expansions- und der Flexionsschicht. Die Steuerungsschicht realisiert den Import der Lexeme des Stammformenlexikons und lässt die eingelesenen Lexeme durch die Expansionsschicht in klassifizierte Vollformen überführen. Über einen XML-Streamer wird das erstellte interne Lexikon als XML-Dokument ausgegeben.

Die Expansionsschicht erhält jeweils eine aus der Stammform generierte Datenstruktur, ein sogenanntes internes Lexem. Daraus ermittelt sie den zugehörigen Paradigmensatz und daraus die jeweiligen Merkmalssätze, die eine Vollform beschreiben (z. B. Gen. Pl.). Der Merkmalssatz wird dann, zusammen mit bestimmten Angaben des internen Lexems, an die Flexionsschicht weitergereicht. Die Merkmalssätze und die erzeugten Vollformen werden gesammelt und an die Steuerungsschicht zurückgegeben.

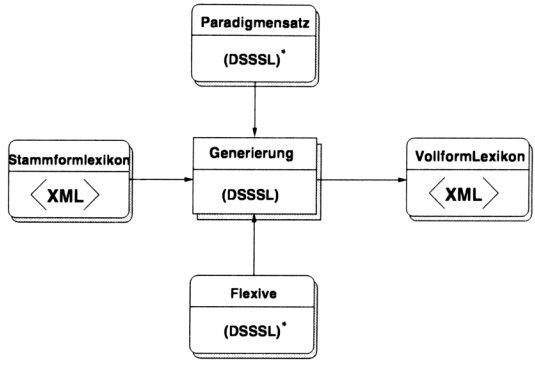

Abbildung 4.1.: Systemarchitektur

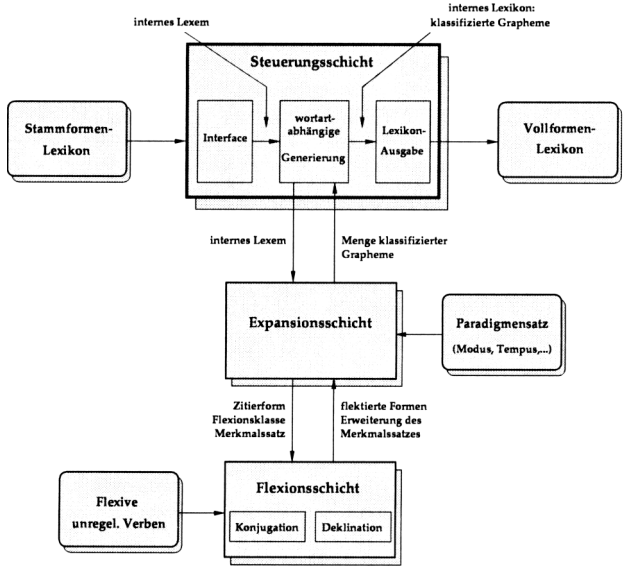

Abbildung 4.2.: Generierung

Die Flexionsschicht ermittelt zu einem gegebenen Lexem und Merkmalssatz auf Grund der Klassennummer bzw. ableitbarer Merkmale des internen Lexems die jeweilige Vollform bzw. Vollformen.

4.3. Stammformlexikon

Die meisten computerlinguistischen Lexika sind primär auf eine effiziente Verarbeitung und Speicherung der Lexikoneinträge ausgerichtet. Daher verwenden sie spezielle Lexikonformate, wie z. B. DATR, zur Repräsentation ihrer Daten. Klassische Wörterbücher hingegen reihen typischerweise die Lexeme in ihren Stammformen aneinander und stellen grammatische, semantische oder auch etymologische Informationen über sie zur Verfügung.

Das hier verwendete Stammformenlexikon basiert auf der Struktur und dem Informationsgehalt von Standardwörterbüchern, wie sie mit dem Duden oder dem Wahrig für das Deutsche vorliegen. Es wird von der Annahme ausgegangen, dass entsprechende Wörterbücher für die Verarbeitung texttechnologisch aufbereitet sind.

Die Generierungskomponente erwartet ein Stammformenlexikon im XML-Format, das in Strukturierung und Annotierung eine eindeutige Interpretation der enthaltenen Informationen ermöglicht. Realistisch an diesem Ansatz ist, dass die Datenhaltung der Wörterbuchverlage (vgl. Kamps et al., 1999) auf SGML basiert. Es ist jedoch unrealistisch, davon auszugehen, diese Daten semantisch markiert vorzufinden obwohl dies auch für andere Anwendungen eine Reihe von Vorteilen mit sich bringt (vgl. Lobin und Witt, 1999). Vielmehr werden in Verlagen üblicherweise optische Unterscheidungsmerkmale (Fettdruck, Kursivschnitt, Symbole etc.) verwendet. Diese müssen daher in einem Vorverarbeitungsprozess eindeutig einer inhaltlichen Funktion zugeordnet werden. Die beiden Formen des Wortes „übersetzen" (aus: Wahrig, 1997) weisen z. B. die erforderliche Eigenschaft der Eindeutigkeit der Markierung auf, um automatisch eine Überführung von visuell ausgerichtetem Markup zu semantischen Auszeichnungen vornehmen zu können:

'über|set·zen <V.i.; ist> ...
über'set·zen <V.t.; hat> ...

Der vertikale Trennstrich (|) markiert ein abtrennbares Präfix , welches durch die Markierung mit einem Apostroph als betont gekennzeichnet ist. Im zweiten Fall ist das „über" durch das Fehlen von Betonungs- und Abtrennbarkeitsmarkierung implizit als nicht abtrennbar und unbetont gekennzeichnet.

Ein Eintrag des Standardlexikons (z. B. 'aus|schal·ten <V.t.; hat> ...) kann in einer hypothetischen, formatorientierten SGML-Datenbasis folgendermaßen vorliegen:

```
<entry>aus&bar;schal&dot;ten
 <spec>&lt;V.tr; hat&gt; </spec>
       ...
</entry>
```

Läßt sich die eindeutige Interpretation der Zeichen etablieren, wie es für gute Wörterbücher gegeben ist, so kann die Datenbasis automatisch in eine semantisch annotierte XML-Repräsentation überführt werden.

```
<lexem id="AUSSCHALTEN">
<graphem>aus<präfix-trennbar/>schal<silbengrenze/>ten</graphem>
  <grammatik>
     <pos>V</pos>
     <trans>tr</trans>
     <aux>hat</aux>
  </grammatik>
  <spezifikation>...</spezifikation>
</lexem>
```

Aus Sicht der Generierung enthält die Mikrostruktur des Stammformenlexikons sowohl relevante Angaben, wie etwa Genus, Klassennummer oder Transitivität als auch irrelevante Angaben, z. B. zur Genealogie oder Semantik des Lexems. Genutzt werden derzeit nur wenige der insgesamt verfügbaren Angaben herkömmlicher Lexika. Durch die Anpassung des Interfaces und der Lexikon-Ausgabe lassen sich jedoch beliebige Informationen in das zu generierende Vollformenlexikon übertragen.

4.4. Vollformenlexikon

In das Vollformenlexikon sollten aus ökonomischen Gründen nur die wortformspezifischen Informationen sowie elementare Angaben, wie die Wortart, aufgenommen werden. Sonstige Angaben können durch einen Rückbezug (Link) auf das Stammformenlexikon (gefahren verweist auf FAHREN) bewahrt werden. Dadurch wird von jeder abgeleiteten graphemischen Form ein Zugriff auf die zugehörigen Lexemangaben ermöglicht. Inwieweit eine direkte Übernahme von lexembezogenen Lexikoninformationen in das graphemorientierte Vollformenlexikon sinnvoller wäre bzw. ob eine Übernahme einzelner Stammformeinträge in das resultierende Vollformenlexikon signifikante Vorteile (lokale Adressierung) ergibt, kann und sollte in Abhängigkeit der die Lexika nutzenden Anwendungen und der Größe des resultierenden Lexikons entschieden werden.

Das Vollformenlexikon enthält die Angaben zu einem Eintrag als Liste von Attribut-Wert-Paaren. Die variablen Angaben, die nicht für jeden Lexikoneintrag existieren, etwa „Numerus = Singular", werden dabei nicht direkt in korrespondierende XML-Attribut-Wert Zuweisungen umgesetzt, sondern in einer generischen Struktur repräsentiert:

```
<Attribut Name='Numerus' Wert='Singular'/>
```

Dies erlaubt die uniforme Behandlung beliebiger Informationseinheiten. Da die Literale, die Name und Wert bezeichnen, nur aus der XML Perspektive atomar sind, nicht notwendigerweise aber aus der Sicht einer auf diesen Daten operierenden Anwendung, lassen sich auch beliebig komplexe Zusammenhänge modellieren.

Ein Beispieleintrag aus dem Vollformlexikon ist nachfolgend zu betrachten:

```
<Eintrag><Graphem>ausgeschaltet</Graphem>
 <Variante>
  <Stamm-Id>ausschalten-V</Stamm-Id>
  <Att Name="Trans" Wert="tr"/>
  <Att Name="Aux" Wert="hat"/>
```

```
  <Att Name="Lemma" Wert="ausschalten"/>
  <Att Name="WA" Wert="VPP"/>
  <Att Name="Temp" Wert="PPerf"/>
 </Variante>
</Eintrag>
```

Typischerweise enthält ein Eintrag mehrere Varianten (Lesarten) eines Graphems (z. B. erste und dritte Person Plural). Das Vollformenlexikon ist somit also einfach strukturiert und nutzt nur wenige der möglichen Restriktionen eines XML-Dokuments. Dadurch wird insbesondere ermöglicht, dass Stammformenlexika mit unterschiedlichsten Angaben eingesetzt werden können.

4.5. Lexemexpansion

Die Expansion der Lexeme wird primär durch die Wortart gesteuert. Grundsätzlich lassen sich flektierende und nicht flektierende Wortarten unterscheiden. Die nicht flektierenden Wortarten müssen nicht weiter betrachtet werden, da die Zitierform die einzig mögliche Wortform ist. Die zusätzlich im Vollformlexikon benötigten Angaben des Stammformlexikons werden durch das Interface in das Vollformenlexikon übernommen.

Die flektierenden Wortarten sind für die interne Verarbeitung in folgende Gruppen aufgeteilt worden:

- konjugierend
  - regelmäßig flektierende Verben (schwache Verben)
  - unregelmäßig flektierende Verben (starke Verben)
- deklinierend
  - Substantive
  - Adjektive
  - unregelmäßig deklinierbare, z. B. Eigennamen, Fremdwörter, ...

Die Gruppe der unregelmäßig deklinierbaren Arten stellt ein konzeptionelles Zugeständnis an die eingeschränkte Mächtigkeit eines geschlossenen Klassensystems und des Ansatzes einer reinen Oberflächenkomposition dar. Sie erlaubt die Zuordnung von Vollformen zu einer Menge von Merkmalssätzen, die für eine Expansion in das Ausgabelexikon herangezogen wird.

In jeder Flexionsklasse sind die Flexive entsprechend der Merkmalskombinationen der Zielvollform repräsentiert. Die Flexionsschicht ermittelt die Flexionsaffixe sowie die nötigen Änderungen, wie etwa Umlautung, um aus der Zitierform eines Lexems die korrespondierende Wortform zu erzeugen.

Die flektierbaren Wortarten werden auf Basis des Wahrig Klassifikationsschemas (Substantive, starke Verben) bzw. der von Darski (1999) beschriebenen Regeln (schwache Verben) behandelt. Das Klassifikationsschema für Substantive musste intern um einige Klassen erweitert werden, da

es stellenweise nicht hinreichend genau für die Anforderungen einer Oberflächenkomposition klassifiziert.

Das Stammformenlexikon wird durch die DSSSL-Verarbeitung eingelesen und anschließend transformiert, konkreter, expandiert. Das lexikalische Wissen stammt somit aus dem Stammformenlexikon, das Wissen über die Regeln ist teilweise algorithmisch, teilweise als zu verarbeitende Datenstrukturen innerhalb der Dokumentsemantik und Programmiersprache DSSSL realisiert. Alternativ hierzu wäre es auch denkbar gewesen, das Regelwissen ebenfalls XML-annotiert vorzuhalten und als zweite Eingabe der Dokumentsemantik zu verwenden.

4.6. Anwendung und Zusammenfassung

Die präsentierte Komponente zur Erzeugung der Vollformen weist folgende Merkmale auf:

- Konsequente Anwendung der semistrukturierten Datenhaltung; sichert Erweiter- sowie Wieder- und Weiterverwendbarkeit der Ergebnisse
- Flexion des Deutschen, ohne Bindung an spezialisierte computerlinguistische Lexika
- Codierung in Attribut-Wert-Matrix-Notation
- Modulare Architektur

In einer Anwendung wurde ein umfangreicher Ausschnitt der Lexeme des von Lehrndorfer (1996) zusammengestellten Lexikons zu einem „Kontrollierten Deutsch" expandiert. Aus diesem Grundwortschatz-Lexikon im Umfang von 1 819 Lexemen wurden 10 396 Grapheme in 42 614 aus der Paradigmenexpansion bedingten Lesarten erzeugt. Diese Lesarten enthalten unter anderem die flexionsbedingten Übereinstimmungen in z. B. Nominativ und Akkusativ einiger Substantive und vor allem der Adjektive, aber auch die „Vollverb"-Anteile diskontinuierlicher Verben. So erzeugen die 227 enthaltenen Adjektive 1 135 orthographische Positiv-Formen in 16 344 Lesarten.

Das Verhältnis von Vollform zu Lesarten ist bei den Verben ca. 1:3, bei den Nomen 1:2,5, und bei den Adjektiven 1:14.

Das erzeugte Vollformenlexikon wurde als Grundlage für einen Tagger verwendet, der Wortinformationen in SGML-Dokumente, speziell in im WWW vorhandene HTML-annotierte Zeitschriftenartikel eingefügt hat.

Literaturverzeichnis

Bird, Steven und Liberman, Mark (1999): "Annotation Graphs as a Framework for Multidimensional Linguistic Data Analysis". In: *Proceedings of the Workshop Towards Standards and Tools for Discourse Tagging*. Association for Computational Linguistics, S. 1–10.

Darski, Józef (1999): *Bildung der Verbformen im Standarddeutschen*. Tübingen: Stauffenburg.

Gippert, Jost (Herausgeber) (1999): *Multilinguale Corpora: Codierung, Strukturierung, Analyse*. Prag: Enigma.

Hausser, Roland (Herausgeber) (1996): *Linguistische Verifikation: Dokumentation zur Ersten Morpholympics 1994*. Niemeyer.

ISO/IEC (1996): "Information Technology – Processing Languages – Document Style Semantics and Specification Language (DSSSL)". Technischer Bericht, ISO/IEC.

Kamps, Thomas; Obermeier, Christoph; Reichenberger, Klaus und Schmidt, Ingrid (1999): *SGML für dynamische Publikationen – das Beispiel Fischer Weltalmanach*, S. 173–192. In: Möhr und Schmidt (1999).

Lehrndorfer, Anne (1996): *Kontrolliertes Deutsch*. Tübingen: Narr.

Lobin, Henning und Witt, Andreas (1999): "Semantic and Thematic Navigation in Electronic Encyclopedias". In: *Redefining the Information Chain – New Ways & Voices. Third ICCC IFIP Conference on Electronic Publishing (EPub '99)*, herausgegeben von Smith, John. ICCC.

Möhr, Wiebke und Schmidt, Ingrid (Herausgeber) (1999): *SGML und XML. Anwendungen und Perspektiven*. Berlin, Heidelberg, New York: Springer.

Pönninghaus, Jens (2000): *Lexikalische Annotierung: Vom Stammformenlexikon zum annotierten Dokument mittels XML und DSSSL*. Diplomarbeit, Technische Fakultät der Universität Bielefeld.

Wahrig (1997): *Wahrig Deutsches Wörterbuch*. Gütersloh: Bertelsmann Lexikon Verlag.

Witt, Andreas (1999): *DSSSL zur Verarbeitung linguistischer Korpora*. In: Gippert (1999).