XML Aufsatz

Wie Wörterbücher mithilfe der Markierungssprache XML Lernern noch mehr Nutzen bringen können

Vorgestellt am Beispiel des elektronischen Wörterbuches WaDokuJT

von Ulrich Apel

1 XML (Einführung)

XML (Extensible Mark-up Language) ist wie HTML Untermenge von SGML; hat jedoch im Gegensatz zu HTML eine Reihe wichtiger Grundgedanken von SGML beibehalten (Erweiterbarkeit, Trennung von Inhalt, Struktur und Formatierung). XML ist seit 1998 ISO-Standard (SGML seit 1986).

SGML: Anfänge Ende der sechziger Jahre bei IBM, als für jede Prozessorengenereration neue Programme geschrieben werden mussten und alte Daten mit jedem neuen Rechner nicht mehr unbedingt kompatibel waren.

SGML insbesondere für sehr umfangreiche Texte z.B. Handbücher zu Flugzeugen, Kraftwerken etc. --> Anpassung an das konkrete Modell, leichte Anpassung von Daten; leichte Wiederverwendung von Daten „Use yesterdays data tomorrow".

Software-unabhängiges Format, Codierung unabhängig von Plattform und Software (Sonderzeichen als Entitäten „ä", „&szlig", „", „&MySign;" etc.), Trennung von Inhalt, Strukturinformationen und Formatierung. Zum Schreiben von SGML genügt ein Texteditor. Bearbeitung der Daten geht per Skript, Datenbankformel, eigenes Programm, Transformation etc.

Problem von SGML – zu kompliziert, zu großer Anfangsaufwand: „Sounds great maybe later". Kein Anwendungsprogramm hat je alle Feinheiten von SGML umgesetzt.

XML reduziert die Komplexität von SGML, und es gibt inzwischen jede Menge Programme die bei der Erstellung, Bearbeitung und Anzeige der Daten helfen (in XML leider keine Möglichkeit mehr Tags abzukürzen).

„Haltbarkeit" der XML-Daten von Millisekunden (z.B. beim automatischen Datenaustausch zwischen Servern) bis zu (hoffentlich) Jahrtausenden (z.B. im Fall der Digitalisierung von Stras).

2 XML und Wörterbücher (Überleitung zu Wörterbüchern)

Codierung nach wie vor ein nicht zu unterschätzendes Problem: Unicode löst nicht alle Probleme, Codierungsstandards aus den 70er und 80er Jahren sind nach wie vor aktuell: Mobiltelefone verwenden Shift-JIS, eingeführte Wörterbuch-Programme EUC, neue Standards haben es schwer. Das muss man bei Datenaustausch und für Rückwärtkompatibilität im Auge behalten.

XML-Code lässt sich bestens in Datenbanken speichern und bearbeiten.

In XML lassen sich zusätzliche Informationen speichern und vor den Nutzern verbergen, die für sie gerade nicht wichtig sind, die aber z.B. für die Bearbeitung, für Kontext etc. wichtig sein können.

Nach entsprechenden Vorgaben, könnte die Anzeige persönlich angepasst werden. Deutsche Muttersprachler könnten z.B. auf Genusangabe verzichten; japanische Muttersprachler brauchen bei Verwendungsbeispielen keine Umschrift etc.

3 Umschrift als Beispiel für die unterschiedlichen Möglichkeit der Repräsentation derselben XML-Daten (konkretes Beispiel für praktische Möglichkeiten)

3.1 Einfaches Beispiel für Umschriften Hepburn- und modifizierte Hepburn-Umschrift

```
<gr>
   <kanji>
      <mora></mora>
      <mora></mora>
   </kanji>
   <kanji>
      <mora></mora>
      <mora></mora>
   </kanji>
<gr>
```

verschiedene Transkriptionsweisen können aus dem XML-Code errechnet werden:

(Kana-Umschrift)

Umrechnungsformel:

--> shi

--> n

--> bu

--> n

shinbun (modifizierte Hepburn-Umschrift)

wenn auf „n" ein „b" folgt --> m

shi_m_bun (Hepburn-Umschrift)

entsprechend zwei phonologische Umschriften

siNbuN

siñbuñ

(0-Akzent --> nicht markiert)

3.2 bessere Beschreibung der Phonetik: Devokalisierung, Nasalierung des stimmhaften velaren Verschlusses, Tonakzent:

XML-Beschreibung der Aussprache:

```
<gr>
   <wdgr>
      <kanji>
         <mora></mora>
         <mora></mora>
      </kanji>
      <kanji>
         <mora dev="yes"></mora>
         <mora></mora>
      </kanji>
   </wdgr>
   <wdgr>
      <kanji>
         <mora  acc="yes"  nas="yes"></mora>
         <mora></mora>
      </kanji>
      <kanji>
         <mora></mora>
      </kanji>
   </wdgr>
</gr>
```

(nur Hiragana z.B. für Nachschlagen der Lesung)

(Hiragana mit Tonakzent z.B. für Lerner, die aus Übungsgründen keine Romaji verwenden wollen; „ka" mit Handaku- ten zur Darstellung der Nasalierung)

kabushiki gaisha (Hepburn-Umschrift)

kabusiki gaisya (Kunrei- shiki, Nihon- shiki)

kabúsiki gàisya (JSL; Japanese: The Spoken Language)

Verschiedene „Geschmacksrichtungen" für phonetische Umschrift.

kabuikiaia (Beginn Hochton: ; Ende Hochton:)

kabikiaia

kabikiaia (Downstep)

etc.

phonetische Umschrift nicht nur zur Darstellung sondern insbes. auch für automatische Sprachgenerierung (diese hatte bislang Probleme mit Phrasierung, Tonakzent und Intonation; Probleme ließen sich mit dem Ansatz abfangen)

Dasselbe Verfahren lässt sich anwenden für „Langvokale" mit Zirkumflex oder mit Macron, Groß- und Kleinschreibung (Satzanfang, Eigennamen) oder für die Markierung von Joshi, um diese gesondert zu behandeln (ha zu wa, bzw. he zu e).

Devokalisierung, Nasaiierung des stimmhaften velaren Verschlusses (bzw. ggf. dessen Fehlen) und Tonakzent sind wichtige Elemente der japanischen Phonetik; werden jedoch in den meisten Wörterbüchern ignoriert (z.B. weil sie sich meistens an japanische Muttersprachler richten).

3.3 Vertiefung Tonakzent (evtl. überspringen):

Japan – „Land der Götter" oder „Land des Papiers"?

Ausrede zu evtl. nicht verfassungskonformem Ausspruch des ehem. Premierminister Mori Yoshiro.

Kami no kuni

:

<kanji>

<mora acc="yes"></mora>

<mora></mora>

</kanji>

:

<kanji>

<mora></mora>

<mora acc="yes"></mora>

</kanji>

(Anm.: Daten liegen nicht in exakt diesem Format vor entsprechende Umformungen und Errechnung des Formats ist möglich)

(Anm.: keine Markierung von „Silben"; eine Definition von Silbe ließe sich jedoch über diese Daten legen)

4. Stand der Markierungen und Ausblick:

4.1 Ist-Zustand

– Genera (ja nach Nutzergruppe ausblenden)

– Sachgebiete (automatische Erstellung von Fachwörterbüchern bzw. Ausblenden von Ortnamen, Persönlichkeiten, Werktitel o.Ä. z.B. weil diese traditionell nicht unbedingt Inhalt eines Wörterbuches sind)

– wissenschaftlicher Name von Tieren, Pflanzen, Begriffen aus Anatomie und Medizin etc.

– Jahreszeitenwörter

– Sprachniveau (z.B. zu „hohes" und zu „flaches" Sprachniveau ausblenden oder z.B. nach Begriffen der Jugendsprache suchen)

– Verwendungsweise (z.B. Onomatopoetika suchen)

– Definitionen (Kompatibiltät zu Wörterbüchern mit Definitionen wie das Chta jisho von Kawamura Yoshiko/Tokyo International University, mit dem wir kooperieren)

– URL, WikiDtsch, WikiJap (Verweis auf Internetseiten außerhalb des Wörterbuches)

– zusätzliche Erklärungen (ggf. ausblenden)

– Familiennamen (wichtig für Suche im Deutschfeld; ließe sich erweitern, dass Namen nach best. Regeln dargestellt werden, z.B. Familienname immer zuerst oder immer zuletzt)

– Herkunft der Einträge (bei Abk. oder Fremdw.; verschiedene Darstellung möglich)

– Beschreibung zur Grammatik

– Markierung zur Funktion von Anführungszeichen (Ironie, wörtliche Rede, Umschrift, Titel, Übersetzung kann ggf. z.B. als Kursiv dargestellt werden)

– häufige Fehler werden ins Wörterbuch aufgenommen, als solche markiert und ggf. ausgeblendet (z.B. „o- share" ; --> Nutzer finden Eintrag, den sie suchen; wenn falscher Eintrag nicht vorhanden ist, kommt er früher oder später sowieso als Nutzereintrag; gilt ähnlich für Fehler aus bestimmten anderen Wörterbüchern)

4.2 weitere Anwendung der bisherigen Markierungen

– Verweise auf Synonyme und Antonyme (könnte/sollte anklickbares Link werden)

– Verweise vom Lemmata auf Ableitungen, Zusammensetzungen, Verwendungsbeispiele und Bespielsätze (mithilfe von XML könnten diese Daten sortiert werden und ihrem Kontext entsprechend angepasst werden, z.B. ausblenden von Redundanz, Ersetzung des Lemmas in Untereintrag durch Tilde).

– es gibt Verweise auf vorhandene fürs Projekt erstellte Bilder. Diese könnten evtl. auch mal tatsächlich angezeigt werden.

4.2 Weitere Möglichkeiten (theoretischer Ausblick):

– Wichtigkeit von einzelnen Übersetzungsvorschlägen, Bedeutungen oder auch Kommentaren markieren; sprich was könnte z.B. Inhalt eines Wörterbuches zum Grundwortschatz sein, was wäre geeignet für ein mittleres Wörterbuch, ein Großwörterbuch oder nur ein Spezialwörterbuch.

Nutzer könnten dann ja nach Schwierigkeitsgrad des Textes wählen, wie sehr das Wörterbuch ins Detail gehen soll.

Entsprechend wäre eine automatische Generierung eines Wörterbuches zum Grundwortschatz bzw. eines mittleres Wörterbuches etc. möglich.

Systeme, die Texte automatisch mit Wortübersetzungen versehen (JGloss, Reading Chta, wwwdic, Popjisyo etc.) benötigen möglichst konzise Einträge. Diese können mit einem solchen Ansatz generiert werden.

– Bessere Markierung von Inhalten, die bei gemeinsamer Darstellung von Lemma und Untereintrag redundant sein könnten.

– Markierung von Paaren aus Lemma und Übersetzung nach der Eignung für ein deutsch-japanisches Wörterbuch (Grundlage für ein solches Wörterbuch)

5 Zusammenfassung

XML gibt Wörterbuch-Daten eine ungeahnte Flexibilität und eröffnet Möglichkeiten, von den man bisher kaum geträumt hat. z.B. Generierung verschiedener Wörterbücher aus denselben Daten oder Anpassung an ganz persönliche Bedürfnisse und Vorlieben.

Erweiterbarkeit: Wörterbuchdaten werden laufend überarbeitet und verbessert. Markierungen werden immer feinstufiger. Kooperation mit und Kompatibilität zu anderen Projekten und Abgleich mit deren Daten helfen inhaltlich voran und erleichtern Korrektur von Fehlern. Nutzer können dadurch auch Zugriff auf deren Daten bekommen z.B. auf die japanischen Wortdefinitionen aus dem Chta-Wörterbuch bzw. deren deutsche Entsprechungen.

Problem: Daten werden sehr komplex. Pflege und Eingabe der Daten ist inhaltlich und formal nicht trivial sondern einigermaßen aufwendig. Einbindung der Nutzer als Mitarbeiter wird dadurch nicht vereinfacht.