

Annotation und Repräsentation morphologischer Strukturen in syllabischen Symbolsystemen¹

In diesem Artikel werden Ansätze vorgestellt, wie komplexe Schriftsysteme kodiert und verschiedene Schriftsysteme zueinander relationiert werden können. Der üblichen Vorgehensweise, für ein Schriftsymbol einen numerischen Identifikator zu definieren, wird die Möglichkeit von Symbolidentifikationen mittels Dokumentauszeichnungen gegenübergestellt. Am Beispiel japanischer morphologischer Strukturen, die in der syllabischen Standardverschriftlichung annotiert werden sollen, wird eine eigene Methodologie entwickelt und implementiert.

Einleitung

Die Standardisierung unterschiedlicher Schriftsysteme und ihrer kulturellen, regionalen, historischen und funktionalen Varianz ist ein prinzipiell nicht abschließbares Unternehmen. Das UNICODE-Konsortium widmet sich dieser Aufgabe und strebt an, die Zeichen aller gebräuchlichen Schriftsysteme der Welt in einem einheitlichen Format zu erfassen. Grundlegendes Prinzip lautet dabei, dass für ein graphisches Zeichen ein singulärer numerischer Identifikator definiert wird. *Graphische* Übereinstimmung von Zeichen ist jedoch nicht in jedem Fall ein hinreichendes Kriterium für die Zeichendeklaration. Birnbaum (1996:206) schreibt in seiner Untersuchung zu Kodierungsverfahren für historische Varianten kyrillischer Schriftsysteme:

„[...] different languages or orthographic systems that use the same script may assign different sounds or meanings to what appear to be the same graphic items.“

Gippert (1999) beschäftigt sich mit der Anwendung von UNICODE für die Kodierung multilingualer Korpora. Problematisch erscheint ihm, dass einige Schriftsysteme nicht Bestandteil der UNICODE-Basisdeklaration aus 16 Bit sind, so dass wenig Software zu ihrer Verarbeitung verfügbar ist. Des Weiteren sind die Schriftrichtung oder verschiedene Zeichenvarianten wie ‚isoliert, initial, zentral, final‘ im Standard als ein Parameter der Zeichenausgabe kodiert und nicht Bestandteil der Zeichendeklaration selbst. Für Schriftsysteme wie das Arabische, die derartigen Varianten bedeutungsdifferenzierendes Gewicht beimessen, ist dieses Vorgehen nicht hinreichend. Problematisch sind etwa diejenigen Fälle, in denen nicht-syllabische Einheiten (Konsonatencluster etc.) in ein an sich syllabisches System integriert werden sollen, ohne typographisch unakzeptable Dokumente hinnehmen zu müssen.

In diesem Artikel wird diese Problematik am Beispiel des Japanischen behandelt, das sich durch eine silbenstrukturierte Schrift in einem komplexen Schriftsystem auszeichnet. Verschriftlichungen, zum Beispiel von morphologischen Strukturen, welche ein zusätzliches, über die syllabischen Zeichen hinausgehendes Symbolinventar benötigen, stellen eine besondere Herausforderung für die Zeichenkodierung und -segmentierung dar, deren hier propagierte Lösung generelle Anwendbarkeit hat.

¹ Dieser Artikel entstand im Rahmen des Projektes „Sekundäre Informationsstrukturierung und vergleichende Diskursanalyse“, Teil der DFG-geförderten Forschergruppe „Texttechnologische Informationsmodellierung“. Siehe hierzu <http://www.text-technology.de>

Japanische Verbmorphologie - Struktur und Annotation

Weite Teile der japanischen Morphologie (Tsuji-mura 1996) lassen sich als agglutinierend einstufen - Morpheme sind eindeutig segmentierbar und besitzen jeweils nur eine Funktion / Bedeutung. Ein Beispiel ist die Verbform *tabesaserarenakatta*, die sich in fünf Morpheme unterteilen lässt:

<i>tabe-</i>	<i>sase-</i>	<i>rare-</i>	<i>na-</i>	<i>katta</i>
essen	KAUSATIV	PASSIV	NEGATION	PRÄTERITUM

Übersetzung: , (Ich) wurde nicht essen gelassen.' , paraphrasiert , (Ich) musste nicht essen.'

Bei der Annotation dieser Strukturen stellt sich das Problem, dass ihre Segmentierbarkeit in der Standardverschriftlichung des Japanischen nicht immer gegeben ist, was im Folgenden exemplifiziert wird. Die Standardverschriftlichung greift auf eine Mischform aus drei Systemen (Coulmas 1996) zurück, die anhand ihrer Gestalt und ihrer prototypischen Verwendungsweise unterschieden werden können. Sogenannte *Kanji*, piktographische Zeichen chinesischen Ursprungs, repräsentieren lexematische Bedeutungen; *Hiragana* sind ein Silbenalphabet und drücken grammatisch-funktionale Einheiten aus wie Kasusmarkierungen oder verbale Suffixe, *Katakana* umfassen das gleiche Silbeninventar (s.u.) wie die *Hiragana* und werden für Lehnwörter aus hauptsächlich westlichen Fremdsprachen benutzt. Das folgende Beispiel gibt einen Satz in der Mischform und die Bestandteile der jeweiligen Schriftsysteme wieder:

Mischform:	私	は	ドイツ	人	です
Kanji	私			人	
Hiragana		は			です
Katakana			ドイツ		
Lateinische Umschrift:	<i>watashi ha</i>		<i>doitsu</i>	<i>jin</i>	<i>desu</i>
	Ich	THEMA	deutsch	Mensch	KOPULA

Übersetzung: , Ich bin Deutscher.'

Die kleinste Einheit der *Hiragana* beziehungsweise *Katakana* bilden sogenannte Mora, die eine Unterklasse von Silben sind, weshalb oft verkürzend von einem Silbenalphabet gesprochen wird. *Nihon* 'Japan' besteht beispielsweise aus zwei Silben *ni-hon*, jedoch aus drei Mora *ni-ho-n*. Durch die beiden Alphabete *Hiragana* und *Katakana* lassen sich alle Mora wiedergeben, die in der sogenannten 50-Laute Tafel wiederzufinden sind:

A	I	U	E	O
Ka	Ki	Ku	Ke	Ko
Sa	Shi	Su	Se	So
Ta	Chi	Tsu	Te	To
Na	Ni	Nu	Ne	No
Ha	Hi	Hu	He	Ho
Ma	Mi	Mu	Me	Mo
Ya	Yu	Yo	Wa	N
Ra	Ri	Ru	Re	Ro

Da mit diesem Schriftsystem die kleinste symbolische Einheit von Silben gebildet wird, lässt sich eine Annotation von Morphemen nur durchführen, wenn ihre Segmentierung kongruent ist zur Silbenebene. Im Bereich japanischer Verben ist dies für die sogenannten vokalischen Verben gewährleistet. Ihr Stamm endet immer auf einen Vokal und ist - wie auch die folgenden Suffixe - mit der obigen 50-Laute-Tafel kongruent. Das Verb *tabe-ru* ‚essen‘ gehört zu dieser Klasse:

Mischschrift: 食ベ る
Lateinisierte, syllabische Schrift *tabe- ru*
Morphemstruktur: tabe- ru
Morphemtransliteration: essen- PRÄSENS

Die konsonantischen Verben lassen sich in Teilen des Formenparadigmas nicht mit dem syllabischen Symbolinventar wiedergeben, da hier der Stamm auf einem Konsonanten endet und somit die Grenzen einzelner Einheiten in der 50-Laute Tafel überschneidet. Das Verb *yom-u* ‚lesen‘ ist ein Vertreter dieser Klasse:

Mischschrift: 読 む
Lateinisierte, syllabische Schrift *yo- mu*

Morphemstruktur: *yom- u*
Morphemtransliteration: lesen- PRÄSENS

Für diese Verben ist die morphologische Segmentierbarkeit in der syllabischen Verschriftlichung also nicht gegeben.

Um dieses Problem zu umgehen, erscheint die Lateinumschrift ein für die Annotation geeignetes Symbolinventar bereit zu stellen, da mit dem lateinischen Alphabet die morphologischen Einheiten hinreichend segmentiert werden können. Kategorien wie Tempus oder Modus eines Verbs, die durch einzelne Morpheme ausgedrückt werden, können unabhängig von den Beschränkungen der gemischten Verschriftlichung segmental annotiert werden. Im folgenden Beispiel wird die Kategorie *VOLITIONAL* an der entsprechenden Stelle segmentiert:

Lateinumschrift:	<i>ik-</i>	<i>ou</i>
Morphemkategorien:	gehen-	VOLITIONAL
Übersetzung:	, lass uns gehen.'	

Für Operationen wie die manuelle Annotation, Suche in Korpora, automatische Lexikonerstellung oder einen die japanische Standardtypographie gewöhnten Benutzer ist die lateinisierte Repräsentation allerdings nicht geeignet, da das Inventar der *Kanji* im Japanischen viele Homophone umfasst. Die Silbenfolge *kou-kou* lässt sich zum Beispiel durch 11 Kombinationen von *Kanji* wiedergeben (hier durch Kommata getrennt), wobei jeder Kombination eine andere lexematische Bedeutung zugeordnet ist:

高校、孝行、航行、口腔、後攻、後項、浩浩、港口、硬膏、皓皓、煌煌

Es stellt sich die Frage, wie die Annotation japanischer morphologischer Strukturen realisiert werden kann, ohne die beschriebenen Vor- und Nachteile des jeweiligen Symbolinventars in Kauf nehmen zu müssen. Zugleich wird damit die generelle Frage berührt, wie mehrere Symbolinventare miteinander in einer Annotation in Beziehung gesetzt werden können. Die einleitend vorgestellte Problematik der Zeichenkodierung wird dabei nicht aus der Perspektive singulärer Zeichen angegangen. Ausgangspunkt bilden vielmehr Strukturierungen von Daten in Dokumenten und ihre Beziehungen zueinander. Eine derartige Vorgehensweise wurde in Gippert (1999) vorgeschlagen, ohne dass konkrete Realisierungsmöglichkeiten erörtert wurden. Dies soll hier versucht werden, unter Berücksichtigung verschiedener Annotationssysteme für das Japanische, der Trennung von Primärdaten und Schriftsystemdeklaration sowie Verfahren zur Relationierung mehrerer Annotationsebenen / Symbolinventare.

Annotations- und Repräsentationssysteme für syllabisch-morphologische Strukturen

Die traditionelle japanische Linguistik (Lewin 1990) hatte keine lateinisierte Umschrift oder eine andere, feinere Symbolinventare zur Verfügung, so dass in ihrem theoretischen Rahmen segmentierende Annotationen morphologischer Einheiten nur begrenzt durchführbar sind. Ist die morphologische Struktur inkongruent zum Symbolinventar der 50-Laute Tafel, so werden die sich überlagernden Einheiten zusammengefasst. Anstelle der obigen Analyse des Verb *yom-u* ‚lesen‘ als aus zwei Symbolen bestehend, d.h. ‚Stamm + Suffix‘, tritt die Analyse als Stamm in der Grundform:

Morphemstruktur-traditionell: `<stem type="basic form">yomu</stem>`

Neuere Annotationssysteme verzichten aus verschiedenen Gründen auf die Granularität der Lateinumschrift und verwenden das syllabische Symbolinventar. Der morphologische Tagger Chasen und mit seiner Hilfe erzeugte Datenbanken morphologischer Muster (Asahara et al. 01)

etwa beruhen auf einer statistischen Analyse der Symbolketten im Primärdatum, also auf die beschriebene Mischform mehrerer Schriftsysteme: die Segmentierung muss auf Grund des Verarbeitungsverfahrens eine syllabische Strukturierung vornehmen. Die im Projekt Verbmobil zur Bildung einer syntaktischen Treebank verwendeten morphologischen Kategorisierungen (Kawata et al. 00) berücksichtigen die morphologische Segmentierung: Es wird zwischen Stamm und Suffix zumindest konzeptuell unterschieden. Die Annotation geht jedoch von der Mischform aus den drei Schriftsystemen aus, so dass nur eine Segmentierung für Stamm und Suffix vorgenommen werden kann. Die konzeptuelle Unterscheidung von Stamm und Suffix, aber die morphologisch nicht adäquat segmentierende Annotation wird auch in anderen Annotationssystemen wie dem LDC-Callhome Korpus (Kobayashi et al. 97) oder in der lexikalisch-morphologischen Datenbank von (Halpern 01) zu Grunde gelegt.

Die TEI (Sperberg-McQueen et al. 94, Kapitel 25) erlaubt es, in einer zu den Primärdaten separaten Form das Schriftsystem in einer ‚writing system declaration‘ (WSD) zu deklarieren. In der WSD können einzelne Sprachen oder auch Subsprachen ausgewählt werden:

```
<language iso639='jpn'>  
Japanese (specialized writing system for waka)  
</language>
```

Soll eine Sprache – wie in unserem Fall – in einer von der Standardverschriftlichung abweichenden Form transkribiert werden, lassen sich entsprechende Transliterationen deklarieren:

```
<character class='lexical'>  
<form string='A' entityStd="agr" ucs-4='03B1' afiiCode='260061'>  
<desc>Greek small letter alpha</desc>  
</form>  
</character>
```

In diesem Beispiel wird der griechische Buchstabe Alpha, im Ausgangsdokument als Lateintransliteration ‚a‘ enthalten, mit der numerischen UNICODE-Deklaration ‚03B1‘ verknüpft, die das native Alpha repräsentiert. Diese Methodologie ist anwendbar, wenn die Repräsentation der japanischen Sillbenschrift transliteriert werden soll in eine lateinisierte Version oder umgekehrt. Dadurch wird jedoch nicht das Problem der unterschiedlichen Segmentierung von Morphemen und Silben gelöst.

Diese Beispiele zeigen, dass die Problematik der Morphemsegmentierung nicht allein durch die Verwendung bestimmter sprachbezogener oder genereller Annotationssysteme zu lösen ist. Die Frage lautet zudem, wie verschiedene Symbolebenen im annotierten Datum miteinander zu verknüpfen sind (siehe hierzu Witt 02, Kapitel 3.2). Potentielle Lösungen bieten zwei Vorgehensweisen, die eigentlich für die Verknüpfung von Annotationsebenen entwickelt wurden. Bird und Liberman gehen in ihrem Ansatz von einer Zeitlinie ‚Timeline‘ aus, welche die kleinste

Segmentierungseinheit bereitstellt und auf die sich alle weiteren Ebenen beziehen lassen. Ausgangspunkt dieser Methodologie (Bird und Liberman 2001) ist eine reale Zeitlinie, die durch eine physikalisch gemessene Verzeitlichung in Form einer Video- oder Audiospur vorliegt. Integrationen dieses Ansatz in ein generelles Format für linguistische Annotationen (Ide et al. 01) eröffnen die Möglichkeit, die Zeitlinie als eine Folge abstrakter Einheiten zu verwenden, mit deren Hilfe das Problem der nicht-kompatiblen Symbolebenen Morphem vs. Silbe sich folgendermaßen lösen lässt (Zur Vereinfachung ist in den folgenden Beispielen die syllabische Annotation in einer Lateinumschrift wiedergegeben):

```

...
<struct type="landmarkDesc">
  <struct type="landmark" id="1"/>
  <struct type="landmark" id="2"/>
  <struct type="landmark" id="3"/>
</struct>
<struct type="syllableAnn">
  <struct type="syllable">
    <startsAt target="#1"/>
    <endsAt target="#1"/>
    <syllable>yo</syllable>
  </struct>
  <struct type="syllable">
    <startsAt target="#2"/>
    <endsAt target="#3"/>
    <syllable>mu</syllable>
  </struct>
</struct>
<struct type="morphemeAnn">
  <struct type="morpheme">
    <startsAt target="#1"/>
    <endsAt target="#2"/>
    <morpheme>yom</morpheme>
  </struct>
  <struct type="morpheme">
    <startsAt target="#3"/>
    <endsAt target="#3"/>
    <morpheme>u</morpheme>
  </struct>
</struct>

```

...

Die zu annotierende Einheit ist wieder das Verb *yom-u* ‚lesen‘. Die Zeitlinie besteht aus drei

Elementen ‚struct‘ mit ‚landmark‘ Attributen, die die feinste Segmentierung *yo-m-u* repräsentieren. Die Elemente ‚struct‘ mit ‚syllableAnn‘ beziehungsweise ‚morphemeAnn‘ Attributen enthalten die Annotationen mit den jeweiligen Symbolinventaren. Mit ‚startsAt‘ und ‚endsAt‘ wird auf die definierten Marken verwiesen. In der Ebene ‚syllable‘ sind dies die Marken ‚#1-#1‘ und ‚#2-#3‘, also die Segmentierung *yo-mu*, in der Ebene ‚morpheme‘ die Marken ‚#1-#2‘ und ‚#3-#3‘, die Segmentierung *yom-u*. Sollen physikalische Zeitmessungen integriert werden, kann dazu in den ‚struct‘-Elementen vom Typ ‚landmark‘ ein ‚position‘ Element eingefügt werden.

Einen anderen Weg geht das Projekt MATE (Pirrelli et al. 00), welches den TEI-Ansatz zur Verknüpfung von Ebenen der Textauszeichnung auf linguistische Annotationen fokussiert. Die grundlegende Symbolebene wird hier mit Identifikatoren ‚ID‘ versehen, auf die in anderen Ebenen referenziert werden kann. Die Ebenen werden in kaskadierender Weise angeordnet, wobei der Bezug zum Primärdatum allein über die Referenzen erhalten bleibt. Orthographische Wörter werden bei MATE als atomare Einheit deklariert, auf die sich die anderen Ebenen beziehen. Es gibt drei Varianten, um die morphologische Struktur auf orthographische Wörter zu beziehen. Entspricht ein lexikalisches Wort einem Morphem, wird mit genau einem Zeiger darauf verwiesen. Soll mehr als ein orthographisches Wort als singuläre morphologische Konstituente gekennzeichnet werden, verweist diese morphologische Einheit auf eine Kette orthographischer Wörter. Der letzte Fall ist derjenige, dass eine morphologische Annotation feiner segmentiert als das orthographische Wort. Dies ist zum Beispiel bei der Annotation von Klitika der Fall. Mit diesen Verfahren lassen sich auch Beziehungen unterschiedlicher Symbolinventare festhalten. Im folgenden Beispiel ist das orthographische Wort die Basisebene, auf die syllabisch und morphologisch referiert wird:

```
...
<w id="w_1">yomu</w>
...
<syll id="syll_1" href="w.xml#id(w_1)>yo</syll>
<syll id="syll_2" href="w.xml#id(w_1)>mu</syll>
<morph id="m_1" type="stem-consonant" href="w.xml#id(w_1)>yom</morph>
<morph id="m_2" type="suffix" href="w.xml#id(w_1)>u</morph>
...
```

Sowohl die Methodik der Zeitlinie als auch die der Referenz auf ein Primärdatum mittels Zeigern sind geeignet, um unterschiedliche Symbolinventare miteinander in Beziehung zu setzen. Beide Verfahren haben den Nachteil, die Beziehungen zwischen den Symbolinventaren nur in den Dokumenten und somit nicht abgeschlossen spezifizieren zu können. Für die nachhaltige Annotation, Analyse und deren Validierung erscheint dies unverzichtbar. Die Zeiger auf Primärdaten haben zudem den Nachteil, dass das Verhältnis von Primärdatum und darauf referenzierender, zugleich aber feiner untergliedernder morphologischer Annotation zu

unbestimmbaren Relationen zwischen den Ebenen führt und keinen Aufschluss gibt über die genaue Segmentierung der Einheiten.

Projektansatz

Um die zwei relevanten Symbolebenen parallel zur Verfügung zu haben und miteinander in Beziehung setzen zu können, werden Informationen über Segmentierungen und lateinbasierte Verschriftlichung² in Attribute eingefügt und ein Tag zur Annotation verwendet, welches die Überlagerung von Morphemen ausdrückt. Anschließend werden Dokumente generiert, welche die notwendigen Segmentierungen in lateinbasierter Form enthalten. Durch weitere Attribute werden Typen von Verben gebildet, die den konsonantischen und vokalischen Verben entsprechen. Für diese Typen gibt es eine abgeschlossene Menge von Regeln, Annotationen einer silbenorientierten Segmentierung in Annotationen einer morphemorientierten Segmentierung zu überführen. Im folgenden wird das Verfahren exemplifiziert.

Das Dokumentfragment DOK-MISCH, eine Annotation in dem Mischsystem, erhält folgende Strukturierung:

```
<stem-suffix romaji_kanji="読む" type="consonant-ending">読む</stem-suffix>  
<stem romaji_kanji="食べ" type="vocal-ending">食べ</stem><suffix romaji_kanji="る">る</suffix>
```

Diese Annotation beruht auf dem gemischten Schriftsystem und muss deshalb für einen Teil der konsonantischen Verben, z.B. *yom-u* ‚lesen‘ auf die Annotations- und Segmentierungseinheit ‚stem-suffix‘ zurückgreifen, um die Überlagerung der morphologischen Einheiten mit dem syllabischen Symbolinventar repräsentieren zu können. Das Verb *tabe-ru* ‚essen‘ als Vertreter der vokalischen Verben hingegen besitzt ein morphologisches Muster, das kongruent zur syllabischen Segmentierung ist. In den Attributen ‚romaji_kanji‘ ist bei beiden Verben die Lateinumschrift und der *Kanji*-Anteil des Datums enthalten. Das Attribute ‚type‘ macht den verbalen Typ explizit, so dass der Prozess der Umwandlung in ein adäquat morphemsegmentiertes Dokument automatisierbar wird. In dem Dokumentfragment DOK-ROM-KANJI ist das ‚romaji_kanji‘-Attribut als Elementinhalt wiedergeben:

```
<stem mix="食べ" type="vocal-ending"> 食べ</stem><suffix mix="る">る</suffix>  
<stem mix="読む" type="consonant-ending"> 読む</stem><suffix mix="0">u</suffix>
```

Dieses Dokument enthält eine Mischform aus Lateinumschrift und *Kanji*, die ausreichend ist für die morphologische Segmentierung. Die Lateinumschrift der *Hiragana* / *Katakana* ist automatisch erzeugbar, die der *Kanji* nicht, da ihre Aussprache wie beschrieben hochgradig ambig ist.

Eine Grundbedingung für die Anwendbarkeit unserer Vorgehensweise ist die Abgeschlossenheit der Regelmenge, welche für die automatische Segmentierung im Dokument DOK-ROM-KANJI

² Standardisierte Dokumentgrammatiken wie XHTML greifen auf eine ähnliche Möglichkeit zurück, Ausspracheinformationen auf der Ebene der Dokumentauszeichnung zu repräsentieren, sog. Ruby-Annotationen.

verwendet wird. Für das Verb *yom-u* ‚lesen‘ handelt es sich um folgendes, hier verkürzt wiedergegebenes Muster:

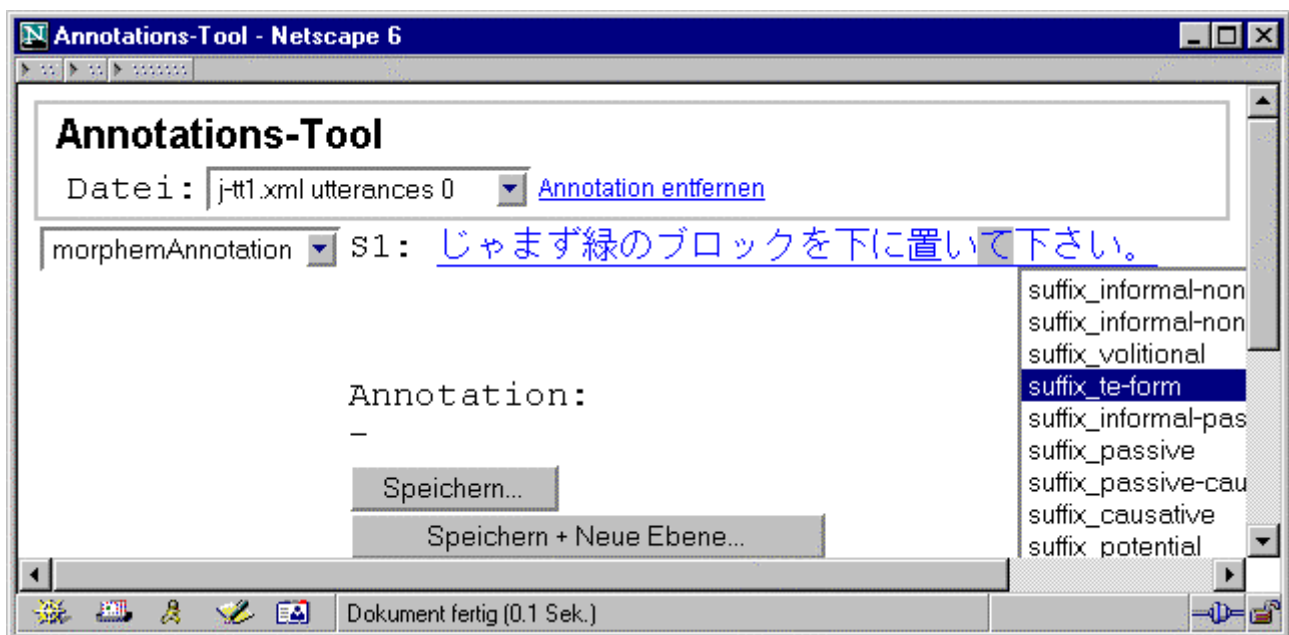
Stamm-Suffix Überlagerung > Neusegmentierung des Dokumentes, Flexionsform des ersten Suffixes (u)

Keine Stamm-Suffix Überlagerung > Keine Neusegmentierung, Überführung in DOK-ROM-KANJI nach dem Muster der vokalischen Verben

Formal eindeutig bestimmt, kann diese Regel im Transformationsprozess zwischen den Dokumenten DOK-ROM-KANJI und DOK-MISCH verwendet werden. Unregelmäßige Verben kennt das Japanische nur im geringen Ausmaß. Die Unregelmäßigkeit betrifft nur das Stammformenparadigma, nicht aber die Agglutination und Segmentierbarkeit der Suffixe, die sich durch weitere Regeln nach dem obigen Muster erfassen lassen.

Implementation: Verarbeitung parallelisierter Symbolebenen

Dieser Abschnitt beschreibt, wie der vorgestellte Ansatz in den Annotations- und Analyseprozess integriert wird. Die Annotation der Primärdaten erfolgt in der gebräuchlichen Mischschrift, aus den beschriebenen Gründen. Verwendet wird für die manuelle Annotation ein webbasiertes Tool:



Der Benutzer kann einzelne Segmente des Datums markieren und ihnen Kategorien zuordnen, etwa 'stem-consonant' oder 'stem-vocal'. Um die Qualität der Annotationen zu sichern, wird sie beim Abspeichern gegenüber einer Dokumentgrammatik validiert. So können unerwünschte Annotationen wie 'suffix+präfix' ausgeschlossen werden.

Als Ausgabe des Tools wird die Datei DOK-MISCH erzeugt und anschließend transformiert in eine Datei DOK-ROM-KANJI. Für den Transformationsprozess wird ein XSL-Stylesheet verwendet, dessen Transformationskomponente XSLT (Clark 99) die Regeln enthält. Der folgende

Ausschnitt aus diesem Stylesheet enthält die obige Regel für die konsonantischen Verben, die in der Mischschrift Stamm-Suffix Überlagerungen enthalten und mit dem Suffix *u* enden:

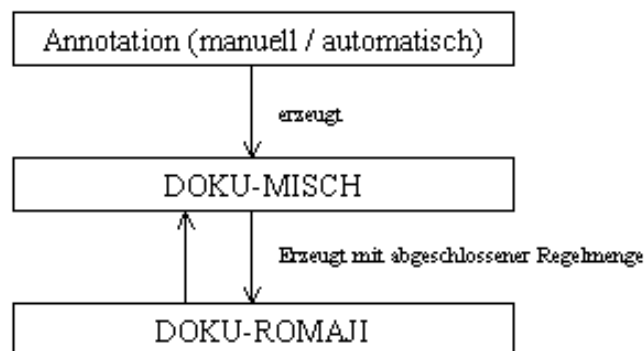
```

<xsl:template match="//stem-suffix[contains('.', 'u')]">
  <stem>
    <xsl:attribute name="mix"><xsl:value-of select="."/></xsl:attribute>
    <xsl:attribute name="type">consonant-ending</xsl:attribute>
    <xsl:value-of select="substring-before('@romaji_kanji', 'u')"/>
  </stem>
  <suffix>
    <xsl:attribute name="mix"><xsl:value-of select="."/>0</xsl:attribute>
    <xsl:text>u</xsl:text>
  </suffix>
</xsl:template>

```

Diese Regel wird auf alle Stamm-Suffix Einheiten angewendet, die auf das Suffix ‚u‘ enden. Aus dem Element ‚stem-suffix‘ wird ein Element ‚stem‘ generiert, das den Inhalt des ‚romaji_kanji‘ - Attributes in den Elementinhalt verwandelt und das ‚type‘ - Attribut beibehält. Zudem wird ein Element ‚suffix‘ generiert, welches das adäquat segmentierte *u* enthält, und ein Attribut ‚mix‘ mit dem Inhalt ‚0‘. ‚0‘ symbolisiert die fehlende Repräsentationsmöglichkeit auf der Silbenebene.

Für weitere Verarbeitungsprozesse, etwa zur Visualisierungen, kann es nötig sein, aus dem Dokument DOK-ROM-KANJI wieder ein Dokument DOK-MISCH zu generieren. Dies geschieht ebenfalls mit einem XSL-Stylesheet. So ist gesichert, dass die Ergebnisse der weiteren Verarbeitung, ebenfalls in die Mischform überführbar sind. Der Datenfluss wird im folgenden Diagramm dargestellt.



Resumee

Der vorliegende Ansatz setzt zwei Symbolinventare durch eine regelbasierte Transformation miteinander in Beziehung und ermöglicht so, auf die Vorteile beider zurückgreifen zu können, ohne ihre Nachteile in Kauf nehmen zu müssen. Dieses Vorgehen wurde am Beispiel der japanischen Verbmorphologie und ihrer verschiedenen Verschriftlichungsformen demonstriert. Die generelle

Problematik der Verschriftlichung komplexer Schriftsysteme beziehungsweise kultureller, regionaler, historischer und funktionaler Schriftvarianten findet in dem propagierten Verfahren eine Lösung, die bestehende Standards für singuläre Symbole wie UNICODE nicht erweitert, sondern die Dokumentauszeichnung als Mittel der Deklaration zusätzlicher Symbolinventare und eventuell notwendiger Segmentierungen verwendet³. Dadurch wird die standardunterstützte Verarbeitung der Daten gesichert, zugleich aber durch die Implementation der Transformation in XSLT eine einfache Anpassung der Methodik auf andere Domänen als die der Morphologie ermöglicht. Grundlegende Voraussetzung ist, dass eine abgeschlossene Regelmenge zur Verfügung steht, die Grundlage der Transformation bildet, und dass die Abbildung der Symbolmengen zueinander eineindeutig ist.

Literatur

- Asahara, M., R. Yoneda und Y. Matsumoto (2001) Use of a relational database in the development and maintenance for statistical Japanese morphological analysis. In: Proceedings of the IRCS workshop on linguistic databases. Philadelphia: University of Pennsylvania.
- Bird, S. und M. Liberman (2001) A formal framework for linguistic annotation. In: Speech and Communication 33 (1,2), 23-60.
- Birnbaum, D. J. (1996) Standardizing characters, glyphs, and SGML entities for encoding early Cyrillic writing. In: Computer Standards and Interfaces 18.
- Clark, J. (1999). XSL Transformations (XSLT) Version 1.0 W3C Recommendation, 16 November 1999.
- Coulmas, F. (1996) Typology of writing systems. In: H. Günther (Hrsg.) Schrift und Schriftlichkeit. Ein internationales Handbuch zeitgenössischer Forschung. Berlin: de Gruyter.
- Dürst, M. und A. Freytag (2002) Unicode in XML and other Markup Languages. Unicode Technical Report #20. W3C Note 18. Februar 2002. <http://www.w3.org/TR/2002/NOTE-unicode-xml-20020218>.
- Gippert, J. (1999) Language-specific encoding in multilingual corpora: requirements and solutions. In: G. J. (Hrsg.) Multilinguale Korpora: Kodierung, Strukturierung, Analyse. Prag: Enigma.
- Halpern, J. (2001) Japanese POS-Codes, morphological attributes and Japanese lexical database. CJK-Institute, <http://www.cjk.org/cjk/samples/jamsam.htm>.
- Ide, N. und L. Romary (2001) Standards for language resources. In: Proceedings of the IRCS workshop on linguistics databases. Philadelphia: University of Pennsylvania.
- Kawata, Y. und J. Bartels (2000) Stylebook for the Japanese treebank in VERBMOBIL. Verbmobil-Report Nr. 240.
- Kobayashi, M., S. Crist, M. Kaneko und C. McLemore (1997) LDC Japanese Lexicon.
- Lewin, B. (1990) Sprachwissenschaft. In: Hammitzsch, H. (Hrsg.) Japanhandbuch. Stuttgart: Franz Steiner Verlag.
- Pirrelli, V. und C. Soria (2000) MATE Dialogue annotation guidelines: Deliverable D2.1, Kapitel „Morphosyntax“. Siehe <http://www.ims.uni-stuttgart.de/projekte/mate/mdag/ms/Intro.htm>.
- Tsujimura, N. (1996) Introduction to Japanese linguistics. Cambridge: Blackwell.
- Sperberg-McQueen, C.M. und L. Burnard (1994) Guidelines for electronic text encoding and interchange (TEIP3). Oxford: Oxford University Computing Services.
- UNICODE. The Unicode Standard. <http://www.unicode.org>.
- Witt, A. (2002) Multiple Informationsstrukturierung mit Auszeichnungssprachen. XML-basierte

³ Zum Verhältnis von UNICODE und Auszeichnungssprachen siehe auch (Dürst et al. 02), insbesondere den Abschnitt „Interlinear annotation characters“.

Methoden und deren Nutzen für die Sprachtechnologie. Dissertation, Universität Bielefeld,
Fakultät für Linguistik und Literaturwissenschaft.