Peter Meyer

Von XML zum DAG

Der lexikographische Prozess bei der Erstellung eines graphenbasierten Wörterbuchportals

1 Einleitung

Die Schaffung übergreifender und einheitlicher Zugriffs- und Recherchemöglichkeiten für bereits vorhandene elektronische Ressourcen unterschiedlicher Herkunft und Struktur ist eine große Herausforderung auch für die gegenwärtige Computerlexikographie. Im Folgenden sollen exemplarisch Lösungsstrategien für Probleme vorgestellt und diskutiert werden, die sich beim Aufbau des *Lehnwortportals Deutsch*, eines unter der URL lwp.ids-mannheim.de frei zugänglichen Internetportals für Lehnwörterbücher, ergeben haben. Der Schwerpunkt soll hierbei auf dem lexikographischen Prozess der Erstellung der wörterbuchübergreifenden Portaldatenbasis liegen. Einleitend wird das Portal vorgestellt und eine kurze Skizze der zugrunde liegenden Datenmodellierung gegeben, die eine XML-Repräsentation der Einzelartikel mit einer graphenbasierten Datenhaltung für wörterbuchübergreifende Suchen verknüpft. Abschnitt 2 diskutiert die verschiedenen Aspekte des lexikographischen Prozesses für das Portal, der in erster Linie die Schaffung von übergreifenden Vernetzungs- und Zugriffsstrukturen beinhaltet, unter dem Gesichtspunkt der technischen sowie inhaltlichen Besonderheiten, die aus einer graphenbasierten Datenhaltung resultieren. Im abschließenden Abschnitt 3 wird die Portalarchitektur allgemein unter dem Gesichtspunkt der Erweiterbarkeit und Wartbarkeit betrachtet.

1.1 Das Lehnwortportal Deutsch

Das am Institut für Deutsche Sprache (IDS) entwickelte *Lehnwortportal Deutsch* (lwp.ids-mannheim.de) ist eine frei zugängliche lexikographische Online-Ressource, die Zugang zu zahlreichen Wörterbüchern für deutsche Lehnwörter in anderen Sprachen ermöglichen soll. Ziel ist es, bestehende sowie speziell für das Portal entwickelte lehnwortlexikographische Ressourcen mit einheitlicher und ressourcenübergreifender Zugriffs- und Vernetzungsstruktur zur Verfügung zu stellen und so den Übergang von der bloßen ‚Translation‘ der Einzelwörterbücher zur ‚Didaktik‘ der Portalpräsentation zu ermöglichen. Neben konventionellen Zugriffsmöglichkeiten auf die Einzelwörterbücher werden daher zum einen wörterbuchübergreifende erweiterte Suchfunktionen angeboten; zum anderen stellt das Portal ein *Herkunfts-*

wörterbuch oder „umgekehrtes Lehnwörterbuch“ (vgl. Engelberg 2010) zur Verfügung, in dem der Weg deutscher Etyma in die verschiedenen Nehmersprachen der integrierten Lehnwörterbücher nachvollzogen werden kann. Die Lemmata des Herkunftswörterbuchs, in der Regel standardsprachliche neuhochdeutsche Wortformen, fungieren als etymologische *Metalemmata* des Portals und werden bei der Portalerstellung definiert. In einem eigenständigen lexikographischen Prozess, der in Abschnitt 2.4 näher beschrieben wird, wird jedes in einem beliebigen Lehnwörterbuch des Portals gebuchte deutsche Etymon (häufig handelt es sich dabei um eine ältere oder dialektale Form) mindestens einem solchen Metalemma zugeordnet. Diese Metalemmata, und nicht etwa die nehmersprachlichen Lemmata der einzelnen Lehnwörterbücher, sind das wichtigste *tertium comparationis* des Lehnwortportals.

Im November 2012 wurde als Ergebnis eines kleineren Pilotprojektes¹ eine erste, noch prototypenartige Fassung des Portals freigeschaltet, die drei bereits früher veröffentlichte Lehnwörterbücher umfasst, nämlich zu deutschen Lehnwörtern im schrift- und standardsprachlichen Polnischen (WDLP; 2447 Wortartikel), im Teschener Dialekt des Polnischen (WDLT; 839 Wortartikel) sowie im Slovenischen (DLS; 1568 Wortartikel ohne reine Verweiseinträge). Die Integration weiterer bereits vorhandener sowie auch einiger dediziert für das Lehnwortportal entwickelter lehnwortlexikographischer Ressourcen ist in Vorbereitung.

Die Zugriffs- und Suchmöglichkeiten des Portals werden in Meyer (2013a) im Einzelnen vorgestellt und in Bezug zur nachfolgend skizzierten Datenmodellierung gesetzt; weitere Informationen hierzu finden sich auf den Dokumentationsseiten des Portals selbst (<http://lwp.ids-mannheim.de/doc/portal/start>). Eine detaillierte Diskussion der graphenbasierten Suchfunktionen findet sich in Meyer (2013b).

Der Screenshot in Abbildung 1 zeigt den Eintrag zum Metalemma *Gestalt* im Herkunftswörterbuch. Er besteht im Wesentlichen aus Links auf zugehörige Einträge in den Lehnwörterbüchern und nennt zusätzlich die in diesen Einträgen genannten diasystematischen Formen des Etymons.

Der Screenshot-Auszug in Abbildung 2 demonstriert am Beispiel des kurzen Eintrags zum slovenischen Lemma *štalt* (entlehnt aus deutsch *Gestalt*) die portalweit einheitliche Präsentation der Einzelwörterbücher mit Alphabetleiste, Lemmalisten-ausschnitt und Schnellsuche mit Auto-Vervollständigung. Die Auswahl des Wörterbuchs geschieht in der rechten Seitenleiste. Zu jedem Artikel werden am Ende der Seite Verweise auf die zugehörigen Einträge (Metalemmata) des umgekehrten Wörterbuchs sowie jeweils eine Auflistung der im Artikel gebuchten zugehörigen Etyma und Lehnwörter samt deren Varianten, Derivaten usw. gegeben.

¹ Das Pilotprojekt wurde vom Beauftragten der Bundesregierung für Kultur und Medien aufgrund eines Beschlusses des Deutschen Bundestages gefördert und im Zeitraum von Mai 2011 bis einschließlich August 2012 durchgeführt.

Institut für
DEUTSCHE SPRACHE

Lehnwortportal
Deutsch

Startseite | Erweiterte Suche | Dokumentation | Bibliografie

Suche nach deutschen Herkunftswörtern

Suchen

Wörterbuch der deutschen Herkunftswörter des Portals - umgekehrtes Lehnwörterbuch
(Institut für Deutsche Sprache)

ABCDEFGHIJKLMNOPQRSTUVWXYZ

Geschäftsmacher
Geschenke
gescherne
Geschichte
Geschirr
Geschirre
Geschlechte
Geschmack
Gesalle
Gesellenstock
Gesenk
Gesenke
Gesicht
Gesimo
gespan
Gestalt
gestalten
Gesapo
Gestell 1
Gestell 2
Gestülbe
Gesuch
Gevatter
Gewächs
Gewähr
gewähren
Gewalt
gewalten 1
gewalten 2
gewaltig
Gewaltiger

Gestalt

Lehnwörterbucheinträge zu diesem Herkunftswort

Polnisch

ksztalt

Gestalt: "Aussehen, Äußeres, Form, Figur einer Person oder..."

gestalt: "Aussehen, Beschaffenheit"

Slovenisch

štalt

gestalt: "Gestalt; Aussehen, Beschaffenheit"

Beachten Sie bitte unsere --Hinweise zu den lexikografischen Beschränkungen dieses Herkunftswörterbuchs.

Sie können eine interaktive graphische Darstellung des Netzwerkes der zu diesem Herkunftswort gehörenden Wortformen und ihrer Relationen zueinander erstellen lassen.

Graphisches Netzwerk erzeugen

Herkunftswörterbuch

Deutsch

Erweiterte Suche

Graphensuche

Dokumentation

Lehnwörterbücher

Polnisch

Tschechener Polnisch

Slovenisch

Abb. 1: Eintrag *Gestalt* im Herkunftswörterbuch des Lehnwortportals

Deutsche Lehnwörter im Slovenischen
(Striedter-Temps 1963)

ABCDEFGHIJKLMNOPQRSTUVWXYZ

→ Zu diesem Artikel gehörige Herkunftswörter

Deutsch

Lehnwörterbücher

Polnisch

Tschechener Polnisch

Slovenisch

Erweiterte Suche

Dokumentation

štal

štal, štāl f. »Gestalt«, k.k.slov. (Ond., Ruge, 1.929); bei Pohl. *štal* dem.; aus mhd. *stalt*, mhd. *gestalt* »Gestalt, Aussehen, Beschaffenheit« (Lex. 1.928); Entlehnung ursprünglich als *štal, dann Schwund des š-. Vgl. Mhd. Fw. 56; RW. 145; Pinner, Let. 1908.178; Bern. RW. 1.635.

Zugehörige Herkunftswörter

Die nachfolgend aufgeführten Einträge des Wörterbuchs der Herkunftswörter des Portals sind diesem Artikel zugeordnet. Klicken Sie auf das >>-Symbol hinter einem Herkunftswort, um die zugehörigen Wortformen aus diesem Artikel zu sehen. Führen Sie den Mauszeiger über eine Wortform, um weitere Information zu ihr zu erhalten.

Gestalt >>

Wortformen vom Typ 'Stylen'

gestalt

Wortformen vom Typ 'Lehnwort'

štalt

Wortformen vom Typ 'Variante'

štavč (Variante zu: Lehnwort)

Abb. 2: Eintrag *štalt* im Wörterbuch zu deutschen Lehnwörtern im Slovenischen

1.2 XML- vs. graphenbasierte Datenmodellierung im Lehnwortportal

Auf der technischen Ebene ist das Lehnwortportal eine am IDS entwickelte Webapplikation, also ein (Java-basiertes) Serverprogramm, das über das Internet herinkommende Anfragen von Browsern mit Hilfe von Daten aus einer relationalen Datenbank beantwortet. Die Datenbank stellt die Einzelartikel der verschiedenen Lehnwörterbücher als XML-Dokumente zur Verfügung, deren Inhalte i. a. über eine sogenannte XSL-Transformation in eine geeignete HTML-Präsentation übersetzt werden. Die Struktur der XML-Dokumente ist nicht einheitlich, sondern unterscheidet sich von Lehnwörterbuch zu Lehnwörterbuch. Für komplexe, insbesondere wörterbuchübergreifende Darstellungen und Suchmöglichkeiten stellt die Datenbank zusätzlich eine aus u. a. den XML-Dokumenten generierte Repräsentationsschicht zur Verfügung, die von den Idiosynkrasien der wörterbuchspezifischen Makro- und Mikrostrukturen sowie Angabeformate abstrahiert. Genauer wird lexikographische Information hier als Menge von z. T. artikel- und wörterbuchübergreifenden *Beziehungen zwischen gebuchten Wortformen*² abgebildet, wobei jeder Wortform diasystematische und grammatische Informationen in einem portalweit einheitlichen Angabeformat zugeordnet sind, vgl. Meyer, Engelberg 2011.

Derzeit werden folgende Arten von Beziehungen zwischen gebuchten Wortformen unterschieden:

(a) Beziehungen von Wortformen eines Lehnwörterbuchartikels zueinander:

- Wortform *x* ist Etymon *y* als *Lehnwort* zugeordnet.
Beispiel: Das polnische Lehnwort *ksztalt* (*x*) ist dem mittelhochdeutschen Etymon *gestalt* (*y*) zugeordnet.
- Etymon/Lehnwort *x* ist (orthographische, phonologische, ...) *Variante* von Etymon/Lehnwort *y*.
Beispiel: Das polnische *kstalt* (*x*) ist phonologische Variante des polnischen Lehnworts *ksztalt* (*y*).
- Metalemma/Etymon/Lehnwort *x* ist morphologische *Ableitung* zu Metalemma/Etymon/Lehnwort *y*.

² Hierbei bezieht sich der im weiteren durchgehend verwendete neutrale Terminus (*gebuchte Wortform*) auf die konkreten, in Wörterbüchern gebuchten Formen von Metalemmata, Etyma und Lehnwörtern, die im Allgemeinen nicht Lemmata der betreffenden Artikel und auch nicht in allen Fällen (z. B. nicht bei orthographischen Varianten) Zitierformen von jeweils unterschiedlichen Lexemen sind.

Beispiel: Das polnische Verb *ksztalcić* (*x*) ist Derivat des polnischen Lehnworts *ksztalt* (*y*).

– Metalemma/Etymon/Lehnwort *x* ist ein *Kompositum*, in dem Metalemma/Etymon/Lehnwort *y* als Kompositionsglied auftaucht.

Beispiel: Das polnische Substantiv *bezksztalt* (*x*) ist Kompositum zum polnischen Lehnwort *ksztalt* (*y*).

– Wortform *x* ist *lexikalische Parallele* zu Lehnwort *y* (derzeit nur für das WDLT benötigt).

Beispiel: Das tschechische Substantiv *drát* (*x*) ist lexikalische Parallele zum Substantiv (Lehnwort) *drót* im Teschener Dialekt des Polnischen (*y*).

(b) Beziehungen zwischen Wortformen in Artikeln verschiedener Wörterbücher:

– Wortform *x* ist Metalemma *y* als (wörterbuchartikelspezifisches) *Etymon* zugeordnet.

Beispiel: Das mittelhochdeutsche Etymon *gestalt* (*x*) ist dem Metalemma *Gestalt* (*y*) zugeordnet.

– Etymon *x* in Artikel *A₂* aus Wörterbuch *W₂* entspricht Lehnwort *y* in Artikel *A₁* in Wörterbuch *W₁* (für Entlehnungsketten benötigt, s. Abschnitt 2.5.).

Beispiel: Das polnische Etymon *drukarz* (*x*) zum ukrainischen Lehnwort *drukar* im derzeit entstehenden Wörterbuch ukrainischer Entlehnungen von polnischen Germanismen entspricht etymologisch dem polnischen Lehnwort *drukarz* (*y*) im Portalwörterbuch zum Standardpolnischen.

Jede gebuchte Wortform im Portal lässt sich als Knoten eines *Graphen* auffassen; die einzelnen eben genannten Beziehungstypen sind entsprechend Relationen, die gerichteten Kanten („Pfeilen“) eines bestimmten Typs zwischen je zwei Knoten (Wortformen) entsprechen. Zwischen zwei gegebenen Knoten (Wortformen) kann es aus inhaltlichen Gründen nur eine Art von Kante (Beziehung) geben. Ohne Beschränkung der Allgemeinheit kann insbesondere aber auch davon ausgegangen werden, dass es, wenn eine Kante $A \rightarrow B$ vorliegt, keine Kante $B \rightarrow A$ desselben Typs geben kann (*Asymmetrie* der Relation): Wenn Wortform *A* Derivat von Wortform *B* ist, kann nicht gleichzeitig *B* auch als Derivat von *A* aufgefasst werden. In vielen Fällen, aber eben nicht z. B. im Falle der Derivation, ergibt sich die „Richtung“ der Beziehung aus anderen Eigenschaften der beiden betroffenen Wortformen, so etwa bei der Relation zwischen Etyma und „zugehörigen“ Lehnwörtern. Eine Sonderfall ist die Beziehung der Variation – zwei Wortformen können einfach als Varianten von-

einander zu werten sein, ohne dass eine Richtung auszuzeichnen ist. In diesem Fall kann man, damit die aus technischen Gründen sehr praktische Annahme der Asymmetrie der Kantenrelationen aufrecht erhalten werden kann, einfach arbiträr eine Kantenrichtung auswählen und ggf. die korrekte semantische Interpretation durch Verwendung eines eigenen Relationstyps (einer besonderen Etikettierung der betreffenden Kanten) sicherstellen (vgl. hierzu 2.2).

Für die folgenden Ausführungen sei festgelegt, dass die Kanten in den oben beschriebenen Relationstypen stets von x nach y gerichtet sind. x wird mithin graphentheoretisch als *Elternknoten*, y als *Kindknoten* aufgefasst. Eine Abfolge von Kanten, die von einem *Vorfahrenknoten* zu einem *Nachfahrenknoten* führt, wird allgemein als *Pfad* im Graphen bezeichnet. Aus inhaltlichen und technischen Gründen gibt es in der für das Lehnwortportal gewählten Modellierung keine *Schleifen*, also keine Pfade, die von einem Knoten zu diesem selbst zurückführen. Es handelt sich insgesamt um einen *gerichteten Graphen ohne Mehrfachkanten und Schleifen*, kurz einen gerichteten azyklischen Graphen (*directed acyclic graph*, abgekürzt DAG). Die erweiterten Suchfunktionen des Lehnwortportals beruhen auf der Suche nach Pfaden in diesem Graphen, vgl. Meyer (2013b).

Gegenüber den ansonsten in linguistischen Kontexten verbreiteten Baumgraphen ist ein DAG eine allgemeinere Datenstruktur, die zulässt, dass ein Knoten mehrere Elternknoten hat. So können einem Lehnwort in einem Wörterbuch durchaus mehrere mögliche Etyma zugeordnet sein; im Fall von mehrfacher Entlehnung können diese verschiedenen Etyma ihrerseits durchaus demselben Metalemma zugeordnet sein. Ein konkretes Beispiel hierzu findet sich am Ende dieses Abschnitts.

Der Graph des Lehnwortportals wird derzeit in einer relationalen Datenbank verwaltet; für die nahe Zukunft ist die Verwendung einer dedizierten Graphendatenbank (Webber et al. 2013) geplant. Jede gebuchte Wortform, also jeder Knoten, wird – mit einer eindeutigen ID-Nummer versehen – zusammen mit weiteren Eigenschaften des Wortes als Datensatz (Zeile) in einer Knotentabelle gespeichert. Da einer Wortform z. B. mehrere Bedeutungsdefinitionen und mehrere grammatische Angaben zugeordnet sein können (sog. 1:n-Beziehungen), werden hierfür eigene Tabellen benötigt, deren Zeilen z. B. jeweils eine Bedeutungsdefinition und die ID des zugehörigen Wortes aus der Wortformentabelle enthalten.

Jede Beziehung (Kante) zwischen zwei Wortformen entspricht einem geordneten Paar der IDs dieser beiden Wortformen und wird entsprechend zusammen mit einer Angabe des Relationstyps als eigene Zeile in eine Kantentabelle eingetragen. Um effizient nach beliebig langen Pfaden suchen zu können, werden auch alle indirekten Nachfahren- und Vorgängerbeziehungen mit entsprechender Etikettierung in der Kantentabelle abgelegt.

Ausschnitte des Graphen – genauer zusammenhängende Teilgraphen, die ein Metalemma mitsamt aller seiner Nachfahren enthalten – können im Herkunftswörterbuch des Portals interaktiv-visuell erkundet werden. Durch bloßes Positionieren des Mauszeigers über einem Knoten werden die zur entsprechenden Wortform ge-

hörenden Attribute angezeigt. Der Screenshot in Abbildung 3 zeigt beispielhaft den im Browser angezeigten Teilgraphen zum Metalemma *Gestalt*. In der Graphenvisualisierung des Portals zeigt die Farbe der als Kreise symbolisierten Knoten deren Zugehörigkeit zu einem Wörterbuch an (Metalemmata grau; WDLP dunkelblau; DLS grün; usw.). Dem Metalemma sind zwei im WDLP genannte Etyma – neuhochdeutsch *Gestalt* mit einer Variante *Gstalt* sowie mittelhochdeutsch *gestalt* – zugeordnet. Es wird Mehrfachentlehnung angenommen, daher sind beide Etyma dem standardpolnischen Lehnwort *kształt* mit der Variante *kształt*, drei Komposita sowie zahlreichen Derivaten zugeordnet. Links oben sieht man das slovenische Lehnwort *štalt* mit der Variante *štavt* sowie dem Etymon *g-stalt*.

Das Wörterbuchnetz des Kompetenzzentrums für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften an der Universität Trier (<http://woerterbuchnetz.de/>) ist ein weiteres Beispiel für ein Wörterbuchportal mit graphenbasierter Datenmodellierung. Hier korrespondieren die Knoten jedoch Wörterbuchartikeln als ganzen, nicht etwa einzelnen in Artikeln gebuchten Wörtern (vgl. Burch/Rapp 2007).

Abb. 3: Lehnwortportal-Teilgraph zum Metalemma *Gestalt*

2 Der lexikographische Prozess zur Erstellung des Wortformengraphen für das *Lehnwortportal Deutsch*

Die Integration von konventionellen digitalen Lehnwörterbüchern in das Lehnwortportal kann nicht sinnvoll automatisiert erfolgen, sondern muss in einem recht komplexen lexikographischen Prozess geleistet werden, der insbesondere die Erzeugung des DAG zum Ziel hat und sich dadurch erheblich von gewöhnlicher lexikographischer Arbeit unterscheidet. In diesem lexikographischen Prozess sind informatische und lexikographische Fragestellungen kaum sinnvoll voneinander zu trennen. Die wesentlichen Aspekte dieses Prozesses werden im Folgenden vorgestellt; die vorgängige Erstellung von konventionellen Lehnwörterbüchern soll hier nicht diskutiert werden.

2.1 Erstellung von XML-Dokumenten

Im Normalfall sollten die lexikographischen Daten der zu integrierenden und zu vernetzenden Lehnwörterbücher bereits in einem inhaltsorientiert ausgezeichneten XML-Format vorliegen. Angesichts des äußerst heterogenen Charakters schon der drei derzeit integrierten Lehnwörterbücher ist die Überführung der Daten in einheitlich strukturierte XML-Dokumente illusorisch bzw. nicht praktikabel. Aus diesem Grund ist die einzige formale Anforderung an die Einzelwörterbücher, dass ihre Artikel jeweils als XML-Dokumente mit einer einheitlichen wörterbuchspezifischen Dokumenttypdefinition (DTD, XML-Schema) kodiert werden können. Sollten solche Dokumente nicht vorliegen, müssen sie i. a. aus den digitalen Originalquellen skriptgesteuert erzeugt werden. Dieser Vorgang, bei dem von Layout- und allgemeiner von Präsentationsaspekten abstrahiert werden sollte, hat eine lexikographische Komponente, ist also nicht rein mechanischer Natur; vgl. die Ausführungen zum sogenannten *lexical view* in den TEI-Richtlinien zum ‚Dictionary‘-Modul (Burnard/Bauman 2007).

Im Fall des Lehnwortportals konnten zwei Wörterbücher (WDLP, WDLT; beide bereits vorher elektronisch publiziert) dank einer Kooperation mit dem Institut für Slavistik der Universität Oldenburg von vornherein in einer geeigneten XML-Modellierung übernommen werden. Die Artikel des Printwörterbuchs DLS, einer lauthistorischen Abhandlung mit Wörterbuchteil, haben diskursiven Charakter und folgen keiner geregelten Mikrostruktur. Für die Online-Darstellung der Einzelartikel des DLS wurde eine Bilddigitalisierung durchgeführt; die Ganzseiten-Scans wurden mit einem eigens entwickelten Softwarewerkzeug manuell in Bilder der Einzelartikel zerlegt und diese den jeweiligen slovenischen Lemmata zugeordnet. Auf den Seiten

des Lehnwortportals wird daher jeder Artikel aus DLS als gescanntes Bild oder, im Falle mehrseitiger Artikel, als Abfolge solcher Bilder dargestellt. Für die Integration in den Graphen war zusätzlich die Erstellung einer XML-Datei pro Artikel erforderlich. Dazu mussten in anspruchsvoller lexikographischer Arbeit wesentliche Angaben zu den im jeweiligen Artikel gebuchten Wortformen in für dieses Wörterbuch geeignet strukturierte Formulare eines Textverarbeitungsprogramms eingetragen werden. In vielen Fällen sind im DLS der genaue Skopus von lexikographischen Angaben oder auch die genauen Verhältnisse zwischen angeführten Wortformen unklar. Die Notwendigkeit der Überführung in ein festes Datenschema erfordert daher an manchen Stellen letztlich arbiträre Entscheidungen und geht häufig mit einer durchaus problematischen Reduktion der Komplexität der Ausgangsartikel einher.

Die Formulare wurden dann in einem Klartextformat exportiert und per Skript in XML-Dokumente fester Struktur überführt. In diesem Fall dienen die XML-Dokumente ausschließlich als Input für die Erzeugung des Graphen; im Normalfall wird aus dem XML jedoch zusätzlich, wie bereits oben erwähnt, mit Hilfe eines wörterbuchspezifischen XSLT-Stylesheets die Online-Ansicht der Artikel erzeugt.

2.2 Extraktion von Teilgraphen aus der XML-Repräsentation I: Graphenstruktur

Aus jedem Artikel eines Lehnwörterbuchs muss für die Portalintegration ein Geflecht von Relationen zwischen Wortformen, genauer ein Teilgraph des portalweiten DAG, 'extrahiert' werden. Falls ein Artikel verschiedene etymologisch nicht zusammengehörige Lehnwörter bespricht, kann es sich sogar um mehrere nicht zusammenhängende Teilgraphen handeln.

Die Erzeugung der Knoten und Kanten des Graphen aufgrund der Informationen in den XML-Dokumenten (Artikeln) eines Wörterbuchs, genauer das Füllen der entsprechenden Datenbanktabellen mit Knoten- und Kanteninformationen, geschieht programmgesteuert. Die Erstellung dieses Programms ist jedoch keine rein informatische, sondern auch eine lexikographische Aufgabe. Da die XML-Dokumente der verschiedenen Lehnwörterbücher sich hinsichtlich ihrer Struktur stark unterscheiden können, muss die Extraktion für jedes Wörterbuch gesondert programmiert werden, denn sogar bei – im Lehnwortportal nicht vorausgesetzter – TEI-konformer Modellierung kann man im Allgemeinen nicht erwarten, Informationen eines bestimmten Typs an strukturell oder durch XML-Auszeichnung eindeutig definierten Positionen des Dokuments zu finden. Die Programmierung ist mit der Klärung inhaltlicher Fragen, insbesondere hinsichtlich der im Graphen anzusetzenden Kanten (Beziehungen), verbunden. Besonders problematisch ist aufgrund der Gerichtetheit des Graphen die Beziehung von (i.d.R. phonetisch-phonologischen) Varianten eines Wortes zueinander. Angenommen, es werden in einem bestimmten Wörterbuch zu einer lemmatisierten Lehnwortform *L1* im zugehörigen Artikel pho-

nologische Varianten *L2*, *L3* und *L4* angeführt; des Weiteren habe auch das zugeordnete Etymon *E1* eine Variante *E2*. Die Verhältnisse zwischen den vier Formen des Lehnworts, falls diese als gleichberechtigt zu gelten haben, ließe sich, etwas unbefriedigend, durch Annahme je einer Variationsbeziehung „Wort *x* ist Variante von Wort *y*“ pro Variantenpaar abbilden, insgesamt also durch sechs Kanten, deren Richtung jeweils arbiträr gewählt werden kann, sofern auf Schleifenfreiheit geachtet wird. Allgemein hätte man bei einem solchen Vorgehen für *n* Varianten insgesamt $n(n+1)/2$ Beziehungen. Alternativ kann man eine Variante, womöglich arbiträr, als die ‚repräsentative‘ herausgreifen, im Beispiel etwa die lemmatisierte Variante *L1*, und die übrigen Formen lediglich als Varianten dieser repräsentativen Form, nicht aber als Varianten voneinander, kennzeichnen. Dadurch werden im Beispiel nur drei, allgemein *n*-1 Beziehungen erzeugt. Analog zu dieser Problematik muss entschieden werden, welche Beziehungen des Typs „Wortform *x* ist Etymon *y* als Lehnwort zugeordnet“ anzusetzen sind: Wird z. B. nur eine einzige derartige Kante, etwa von *E1* nach *L1*, angenommen oder wird *E1* und ggf. auch *E2* mit jeder vorhandenen Lehnwortvariante in Beziehung gesetzt (wodurch bis zu acht Kanten entstehen)? Abbildung 4 illustriert die beiden möglichen Extremfälle; die durchgezogenen Pfeile (Kanten) symbolisieren die Beziehung von Varianten zueinander, die gestrichelten Pfeile verknüpfen Etyma mit zugehörigen Lehnwörtern. Offensichtlich gibt es hier keine pauschal richtige oder falsche Vorgehensweise, zumal die korrekte Relationierung von gebuchten Wortformen im Ausgangsartikel je nach Mikrostruktur des betrachteten Wörterbuchs möglicherweise implizit bleibt und es eines philologisch geschulten Lesers bedarf, um herauszufinden, dass etwa Variante *E1* den Lehnwortformen *L1* und *L2* zugrunde liegt, während *L3* und *L4* auf die Form *E2* zurückgehen.

Abb. 4: Beispiel für minimale (a) und maximale (b) Vernetzung von zwei Etymon- und vier zugehörigen Lehnwortvarianten im gerichteten azyklischen Graphen

Bei einer automatisierten Extraktion von Grapheninformationen gehen in solchen Fällen implizite oder nur in einem diskursiven Kommentar oder anderweitig unsystematisch behandelte Unterscheidungen dieser Art verloren.

Die Wahl der Graphenstruktur lässt sich zudem nicht sinnvoll von der Entscheidung zwischen verschiedenen möglichen Algorithmen für Suchen im Graphen trennen. Im HTML-Formular für erweiterte Suchen im Lehnwortportal lassen sich Suchkriterien sowohl für Etyma als auch für Lehnwörter eingeben; das System sucht dann nach im Graphen nach *Paaren* aus je einem Etymon und einem ‚zugehörigen‘ Lehnwort, bei denen beide Wörter ihre jeweiligen Suchkriterien erfüllen. Möchte man, dass in der Graphenstruktur aus Abbildung 4(a) auch z. B. Lehnwort *L2* als zu Etymon *E2* ‚zugehörig‘ zu interpretieren ist, muss man im Graphen allgemein nach passenden Paaren aus einem Etymonknoten *E* und einem Lehnwortknoten *L* suchen, die einen Etymonknoten *E** als *gemeinsamen Vorgänger* haben, wobei *E** auch identisch mit *E* sein darf. So ist in Abbildung 4(a) Knoten *E1* der gemeinsame Vorgänger von *E2* und *L2*. Bei einer maximalen Vernetzung wie in Abbildung 4(b) ist das formale Kriterium für ‚Zugehörigkeit‘ eines Lehnworts zu einem Etymon einfacher zu formulieren: Es muss ein Pfad vom Etymon zum Lehnwort führen.

Im Lehnwortportal wird derzeit für alle drei integrierten Lehnwörterbücher eine minimalistische Graphenstruktur vom Typ 4(a) angenommen, wobei die Auswahl der ‚Hauptvariante‘ unter den Etyma bzw. Lehnwörtern in einigen Fällen mechanisch und daher in gewisser Weise arbiträr vorgenommen wurde; Kriterium für etymologische ‚Zugehörigkeit‘ eines Lehnworts zu einem Etymon ist die Existenz eines Metalemmaknotens als gemeinsamen Vorgängers beider Knoten. Diese Entscheidung beruht wesentlich darauf, dass die Mikrostrukturen aller drei Wörterbücher keine formalisierte Zuordnung von Etymon- zu Lehnwortvarianten leisten und aus grundsätzlichen philologischen Gründen auch kaum leisten könnten. Sollte jedoch zu einem späteren Zeitpunkt ein Wörterbuch integriert werden, das eine formalisierte Zuordnung vorsieht, könnte dies eine Restrukturierung des Graphen erforderlich machen, die für die jetzt bereits eingegliederten Wörterbücher eine ‚maximale‘ Modellierung vom Typ 4(b) vorsieht und als graphentheoretisches Äquivalent der etymologischen ‚Zugehörigkeit‘ nur noch die Existenz eines direkten Pfades vom Etymon zum Lehnwort vorsieht.

Wenn eine Entscheidung für die zu einem Wörterbuch gehörige Graphenmodellierung gefallen ist, kann auf der Grundlage der XML-Struktur der Artikel des Wörterbuchs die Erstellung der zu den Artikeln gehörenden Teilgraphen programmiert werden. Das Programm muss je bestimmte strukturelle Verhältnisse im XML-Dokumentenbaum in Graphenkanten ‚übersetzen‘. Beispielshalber sei in der folgenden Abbildung 5 ein schematisches XML-Dokument für einen Lehnwörterbuchartikel mit Neststruktur betrachtet; eine solche Struktur findet sich häufig im DLS, wenn z. B. ein slovenisches Substantiv mitsamt einem zugehörigen denominalen Verb verschiedenen deutschen Etyma mit demselben Wortbildungsverhältnis zugeordnet werden, z. B. slovenisch *antlati* und (*h*)*antel* zu deutsch *handeln* bzw. *Handel*.

Auch bei sehr komplexen Dokumenttypdefinitionen reichen i. a. zwei Ausdrücke der XML-Abfragesprache XPath (vgl. <http://www.w3.org/TR/xpath20/> für die jeweils aktuelle Fassung der XPath-Spezifikation in der derzeitigen Version 2.0) aus, um eine in Graphenkanten zu ‚übersetzende‘ strukturelle Relation im XML-Dokument zu spezifizieren: Der erste XPath-Ausdruck adressiert mit einem absoluten Pfad diejenigen XML-Elemente, die Elternknoten der im Graphen einzufügenden Kanten korrespondieren. Im Beispieldokument entsprechen etwa die Nestlemmata eines Artikels XML-Elementen mit dem Namen **Lehnwort**, deren strukturelle Position durch den XPath-Ausdruck `//Lehnwort` (oder expliziter `/Artikel/Lemma/Lehnwort`) beschrieben wird. Der zweite XPath-Ausdruck adressiert dann relativ zu den so bestimmten XML-Elementen das XML-Äquivalent des/der jeweils zugehörigen Kindknoten(s) der zu extrahierenden Kanten. Varianten eines Lehnwortes werden im Beispiel durch XML-Elemente mit dem zum Elternelement relativen Pfad **Variante/Variante** oder kürzer `../Variante` adressiert. Bei einer Modellierung entsprechend Abbildung 2(a) könnte man pro Nestlemma jeweils dem ersten aufgeführten Etymon-Element (absolute Pfadangabe `/Artikel/Lemma/Etyma/Etymon [1]`) als ‚Hauptform‘ die übrigen Etymon-Elemente mit demselben Elternknoten (relative Pfadangabe `../Etymon [position()>1]`) als dessen Varianten zuordnen. Und schließlich gehört zu jeder Etymon-Hauptform `//Etymon[1]` eine lemmatisierte Lehnwortform `../Lehnwort`.

```
<?xml version="1.0" encoding="UTF-8"?>
<Artikel>
  <Lemma>
    <Lehnwort>
      <Wortform>antel</Wortform>
      <Grammatik>...</Grammatik>
      <Varianten>
        <Variante>
          <Wortform>hantel</Wortform>
          <Grammatik>...</Grammatik>
          ...
        </Variante>
      </Varianten>
    </Lehnwort>
    <Etyma>
      <Etymon>
        <Wortform>hantel</Wortform>
        <Grammatik>...</Grammatik>
      </Etymon>
      <Etymon>
        <Wortform>hantel</Wortform>
        <Grammatik>...</Grammatik>
      </Etymon>
    </Etyma>
  </Lemma>
  <Lemma>
    ...
  </Lemma>
</Artikel>
```

Abb. 5: Schematisches Beispiel für das XML-Dokument zum Wörterbuch-Nesteintrag für das slovenisch *antel*

2.3 Extraktion von Teilgraphen aus der XML-Repräsentation II: Attribute von Knoten

Für portalweite Suchvorgänge müssen die womöglich heterogenen und unterschiedlich granularen Angaben der Ausgangswörterbücher insbesondere zu diasystematischen Zuordnungen von gebuchten Wortformen (Sprachstufen bzw. Dialekten) sowie zum Entlehnungszeitraum und zu grammatischen Eigenschaften in ein einheitliches portalweit gültiges Vokabular übersetzt werden. Die zugehörigen konzeptuellen Entscheidungen sind natürlich wiederum genuin linguistischer und lexikographischer Natur. Insbesondere im Falle von raumzeitlichen bzw. diasystematischen Angaben muss eine entsprechende Begriffstaxonomie bzw. eine formale Ontologie festgelegt werden. Im Lehnwortportal wird etwa zur Zeit eine einfache hierarchische Taxonomie von deutschen Dialekten verwendet, die in etwa dem in der heutigen Germanistik üblichen Stand entsprechen sollte (vgl. Wiesinger 1983). So gibt es eine Dialektkategorie *schlesisch*, die Element der Kategorien *ostmitteldeutsch* und weiter *mitteldeutsch*, *hochdeutsch* und *deutsch* ist. Für eine in einem Wörterbuch als schlesisch klassifizierte Wortform *F* wird der zugehörige Graphenknoten mit *allen* vorgenannten Kategorien (Attributen) annotiert, so dass beispielsweise bei einer Suche nach hochdeutschen oder mitteldeutschen Etyma eben auch die Wortform *F* berücksichtigt wird. Die Berücksichtigung von weiteren Informationen, beispielsweise von konkret in Jahres- oder Jahrhundertzahlen spezifizierten Entlehnungszeiträumen, ist wesentlich aufwändiger und würde eine formale Ontologie von Zeiträumen erfordern.

Natürlich hat jede Form der Überführung in ein ressourcenübergreifend einheitliches Datenformat für lexikographische Angaben mit prinzipiellen Problemen zu kämpfen. So kann nicht davon ausgegangen werden, dass verschiedene Lehnwörterbücher für Sprachstufenangaben wie *mittelhochdeutsch* einheitliche zeitliche Grenzen voraussetzen; üblicherweise fehlt sogar jeglicher Hinweis auf die im jeweiligen Werk unterstellten Periodisierungen.

2.4 Erstellung der Metalemmaliste

Die mit Abstand komplexeste lexikographische Aufgabe ist die der Zuordnung von wörterbuchspezifischen Etymon-Wortformen zu ‚normalisierten‘ Metalemmata, idealerweise heute gebräuchlichen standardsprachlichen Wörtern des Deutschen; vgl. hierzu ausführlicher Meyer (2013b). Diese Zuordnung schafft, wie aus Abbildung 3 beispielhaft ersichtlich ist, die wörterbuchübergreifende Verbindung zwischen den aus einzelnen Artikeln extrahierten Teilgraphen, indem verschiedene etymologisch zusammengehörige, typischerweise diasystematisch variierende, Formen eines Etymons mit einem und demselben Metalemma verknüpft und so abs-

trakt miteinander identifiziert werden. Zugleich dienen die Metalemmata, wie bereits erwähnt, als Lemmata des ‚umgekehrten Lehnwörterbuchs‘ (Herkunftswörterbuchs) des Lehnwortportals; jeder Lehnwörterbuchartikel enthält Verweise auf den oder die zugehörigen Herkunftswörterbuchartikel und umgekehrt. Das Herkunftswörterbuch ist also nicht etwa, wie dies in dem monumentalen Wörterbuch niederländischer Wörter in den Sprachen der Welt von van der Sijs (2010) der Fall ist, eine völlig eigenständige lexikographische Ressource, sondern wird durch manuelle Vernetzung der im Portal eingebundenen Lehnwörterbücher sowie durch Anreicherung mit zusätzlicher Information – der Gestalt des Metalemmas – generiert. Als einziges Wörterbuch des Portals haben die Artikel des Herkunftswörterbuchs keinerlei XML-Repräsentation; die Online-Ansicht der Artikel wird dynamisch direkt aus dem Graphen generiert.

Die Erstellung der Metalemmaliste ist nur mit hochspezialisierten Werkzeugen sinnvoll möglich, da für die lexikographische Arbeit sämtliche in das Portal integrierten Lehnwörterbücher simultan verfügbar sein müssen. Am Institut für Deutsche Sprache wurde für die Erstversion des Lehnwortportals eine eigene Redaktionsumgebung für diese Aufgabe entwickelt, die sämtliche Artikel der (derzeit drei) zu integrierenden Wörterbücher als XML-Dokumente einliest und mithilfe wörterbuchspezifischer XPath-Adressierungen der Angaben zu Etyma eine alphabetisch geordnete Liste aller in den XML-Dokumenten gebuchten Etymonformen erstellt. Dieser Liste ist in der Benutzeroberfläche der Redaktionsanwendung eine zweite, anfängliche leere Liste der Metalemmata gegenübergestellt. Neue Metalemmata werden vom Lexikographen einfach durch Klicken auf eine Schaltfläche und anschließende Eingabe eines deutschen Wortes erzeugt. Durch Auswählen eines Etymons aus der einen sowie eines Metalemmas aus der anderen Liste und anschließendes Klicken auf eine Schaltfläche wird eine Beziehung (Graphenkante) vom Metalemma zum Etymon hergestellt. Das Programm ermöglicht auch die Erfassung beispielsweise von Derivations- und Kompositionsbeziehungen zwischen Metalemmata. Sowohl die Etymon- als auch die Metalemmaliste haben schon jetzt mehrere tausend Einträge, so dass für effizientes Arbeiten schnelles und unscharfes Suchen auf den beiden genannten Listen möglich sein muss. Wesentliche Aufgabe der Redaktionsumgebung ist die Automatisierung komplexer Bearbeitungsschritte. Hier sei insbesondere das schon erwähnte Anlegen von Beziehungen zwischen Metalemmata, das nachträgliche ‚Zusammenlegen‘ zweier Metalemmata sowie das ‚Umbenennen‘ und Löschen von Metalemmata genannt. Bei diesen Aufgaben muss jeweils das gesamte Beziehungsgeflecht der betreffenden Metalemmata mitberücksichtigt und Datenkonsistenz gewahrt werden. Beim Erstellen neuer Beziehungen muss insbesondere stets geprüft werden, ob nicht durch Eingabefehler eine Schleife im Graphen entstanden ist.

Die Metalemma-Redaktionsumgebung speichert die Liste der Metalemmata, jeweils mit exakt adressierten Verweisen auf die ihnen zugeordneten Etyma aus den einzelnen Wörterbuchartikeln, einschließlich der zwischen den Metalemmata ange-

legten Kanten sowie diverser weiterer Informationen (Eingabekommentare des Bearbeiters; „todo“-Flagge) in einer recht einfach strukturierten XML-Datei ab. Diese Datei wird beim Generieren des Graphen eingelesen und zur Erzeugung der Kanten zwischen Metalemmata und Etyma genutzt.

Neben der technischen Komplexität dieses Vorgangs stehen die beträchtlichen philologisch-lexikographischen Schwierigkeiten bei der Erstellung der Metalemma-liste. Häufig haben die in den Lehnwörterbüchern genannten Etyma älterer Sprachstufen beispielsweise überhaupt keine oder nur eine dialektale oder morphologisch nicht hinreichend genau übereinstimmende neuhochdeutsche Fortsetzung. Auch die etymologische Identifizierbarkeit von ähnlichen Etymon-Wortformen aus verschiedenen Wörterbüchern ist häufig fraglich.

2.5 Erfassung von Lehnwortketten

Einer der größten Vorteile einer graphenbasierten Repräsentation der Portaldaten ist die Möglichkeit, beliebig lange und komplexe Ketten von Entlehnungsprozessen direkt in einer dafür geeigneten Datenstruktur zu erfassen und dadurch gut präsentierbar und durchsuchbar zu machen. Die derzeitige Version des Lehnwortportals erfasst noch keine Entlehnungsketten. Im Rahmen einer Kooperation des IDS mit dem Institut für Slavistik der Universität Oldenburg wird jedoch derzeit in einem von der Deutschen Forschungsgemeinschaft geförderten Projekt der Weg der im WDLP angeführten deutschen Lehnwörter des Standardpolnischen in die ostslavischen Sprachen Russisch, Weißrussisch und Ukrainisch auf der Grundlage umfangreichen publizierten bzw. zumindest in Karteien erfassten lexikographischen Materials untersucht. Ziel ist die Erweiterung des Lehnwortportals um drei neue Lehnwörterbücher, die im Wesentlichen die polnisch vermittelten ostslavischen Entlehnungen aus dem Deutschen dokumentieren. Zu den einzelsprachlichen Artikeln der drei neuen Lehnwörterbücher wird es jeweils einen sprachübergreifenden Freitextkommentar zu etymologisch zusammengehörigen ostslavischen Lehnwörtern geben, der u. a. die oft komplexe und nicht endgültig aufklärbare Geschichte der Entlehnungsprozesse diskutiert. So ist häufig anstelle oder auch neben der polnisch vermittelten Entlehnung in eine ostslavische Sprache auch die direkte Entlehnung aus dem Deutschen mit dem bekannten Material vereinbar; auch eine z. B. über das Polnische und weiter über das Ukrainische vermittelte Entlehnung ins Russische ist in manchen Fällen plausibel.

Die lexikographische Bearbeitung der Einzelartikel muss aufgrund der technischen Komplexität der zu verwaltenden Vernetzungsinformationen wiederum in einer eigens entwickelten netzwerk- und mehrbenutzerfähigen Redaktionsumgebung erfolgen, die zum einen die bequeme Bezugnahme auf Buchungen des WDLP und zum anderen die synoptische Erstellung mehrerer zusammengehörender einzelsprachlicher Wörterbuchartikel ermöglicht. Im Zuge der Erstellung des ostsla-

visch-sprachübergreifenden Freitextkommentars können die Bearbeiter in dieser Umgebung in einer Graphik durch Auswählen von als Pfeil dargestellten Beziehungen zwischen den fünf beteiligten Sprachen mögliche Entlehnungswege spezifizieren und auch mit einer Plausibilitätsbewertung versehen. Jeder so eingegebene potentielle Entlehnungsprozess wird jeweils im von der Redaktionsumgebung erzeugten XML-Dokument zum betreffenden ostslavischen Lehnwort mit einer exakten Adressierung der zugehörigen Etymon-Wortform aus dem XML-Dokument der Gegersprache kodiert. Bei der Generierung des DAG kann diese Information direkt in *gewichtete* Kanten übersetzt werden; die Gewichtung kann wiederum in eine visuelle Darstellung möglicher Entlehnungswege innerhalb der ostslavisch-sprachübergreifenden Freitextkommentare einfließen und ggf. auch für eine Sortierung von Suchergebnissen nach Plausibilität genutzt werden.

Als Resultat des hier nur grob angedeuteten lexikographischen Prozesses erscheinen mögliche Entlehnungswege im elementaren Fall als Pfade von einem deutschen Etymon über ein polnisches hin zu einem ostslavischen Lehnwort, z. B. *Drucker* (deutsches Etymon) → *drukarsz* (polnisches Lehnwort) → *drukarsz* (polnisches Etymon) → *drukars* (ukrainisches Lehnwort). Das vermittelnde polnische Wort erscheint auf dem Pfad scheinbar doppelt. Im Allgemeinen wird nämlich ein ‚Mitteglied‘ in einer Entlehnungskette in zwei verschiedenen Wörterbüchern auftauchen, einmal z. B. als Lehnwort aus dem Deutschen, einmal als Etymon für ein Wort in einer sekundären Nehmersprache. Die Identifikation der beiden Wortformen, zu denen durchaus divergierende Information in den beiden beteiligten Wörterbüchern zu finden sein mag, ist dann ein eigenständiger Schritt im lexikographischen Prozess der Graphenerstellung, der eine Redaktionsumgebung ähnlich der für die Metalemmaliste erforderlich macht. – Die Besonderheit im vorliegenden Fall ist jedoch, dass die neu einzubindenden lexikographischen Ressourcen dediziert für das Lehnwortportal erstellt werden und der beschriebene Schritt der nachträglichen Identifikation entfällt. Dennoch muss die Graphenmodellierung einheitlich erfolgen.

3 Zusammenfassung und Ausblick

Das *Lehnwortportal Deutsch* des IDS verfügt über eine modulare Architektur, die es ermöglicht, mit geringem technischen Aufwand digitale lehnwortlexikographische Ressourcen unterschiedlichster Struktur und Zielsetzung miteinander zu vernetzen und mit wörterbuchübergreifenden Zugriffsstrukturen zu versehen, ohne dabei die Integrität der Originalwörterbücher zu verletzen. Wesentliche Aspekte der Mikro- und Mediostruktur der Wörterbücher werden artikel- und ressourcenübergreifend durch eine abstrahierende, graphenbasierte Datenhaltungsschicht modelliert. Diese Modellierung ist, dies sollten die vorangehenden Ausführungen zeigen, kein me-

chanischer Vorgang – und das Lehnwortportal mithin kein bloßes ‚Mashup‘ von digitalen Ressourcen –, sondern das Resultat eines vielschichtigen lexikographischen Prozesses, dessen Abläufe und Besonderheiten sich aus der verwendeten Datenstruktur ergeben. Zu diesen Abläufen gehören, wie gesehen, auch Implementierungsarbeiten insbesondere für die automatisierte Extraktion von Teilgraphen aus XML-Dokumenten sowie für die Entwicklung von Vernetzungswerkzeugen für Metalemmata und Entlehnungsketten. Dabei handelt es sich jedoch insofern nicht um rein technische oder informatische Arbeiten, als genuin lexikographische Entscheidungen in die Programmierung einfließen müssen.

Die verschiedenen XML-basierten Datenquellen, die der Erzeugung des gerichteten azyklischen Wortformgraphen zugrunde liegen – Einzelwörterbücher, Metalemmaliste sowie Vernetzungsinformationen für Entlehnungsketten – werden in der Architektur des Lehnwortportals getrennt vom Graphen verwaltet und können im Prinzip jederzeit einzeln und unabhängig voneinander verändert werden. Werden Modifikationen an den Quellen vorgenommen oder kommt z. B. ein neues Wörterbuch hinzu, muss der Graph vollständig neu erzeugt werden. Ein großer Vorteil dieses Vorgehens ist, dass jederzeit weitreichende Änderungen an der grundsätzlichen Graphenstruktur vorgenommen werden können; dies betrifft beispielsweise die oben in 2.2. diskutierte Entscheidung für die Vernetzungsstruktur und die in 2.3. diskutierte Wahl übergreifender Datenformate und Ontologien. Der Graph kann jederzeit um neue Typen von portalspezifischen Informationen angereichert werden, beispielsweise um morphologische oder phonologische Analysen der deutschen Etyma oder um semantische Informationen.

Eine zentrales und für das Lehnwortportal noch nicht gelöstes Problem der Gesamtarchitektur stellen die Interdependenzen dar, die sich durch die Vernetzung mit Metalemmata und die Erstellung von Entlehnungsketten ergeben: Werden Artikel in einem Lehnwörterbuch nach erfolgter Portalintegration überarbeitet, müssen zugehörige wörterbuchübergreifende Vernetzungen überprüft und ggf. angepasst werden. Dabei können Metalemmata wegfallen, neue hinzukommen, Entlehnungsketten verändert werden und vieles mehr. Dies erfordert eine systematische datenbankseitige Abbildung sämtlicher erforderlichen lexikographischen Arbeiten und der dabei verwendeten und erzeugten lexikographischen Informationen sowie eine weitestgehend automatisierte Konsistenzprüfung.

Literatur

(a) Wörterbücher

- WDLP = WÖRTERBUCH DER DEUTSCHEN LEHNWÖRTER IN DER POLNISCHEN SCHRIFT- UND STANDARDSPRACHE. VON DEN ANFÄNGEN DES POLNISCHEN SCHRIFTTUMS BIS IN DIE MITTE DES 20. JAHRHUNDERTS. Hgg. Andrzej de Vincenz, Gerd Hentschel. Oldenburg: BIS-Verlag 2010 (= Studia Slavica Oldenburgensia 20). (<http://diglib.bis.uni-oldenburg.de/bis-verlag/wdlp>) (07.11.2013)
- WDLT = WÖRTERBUCH DER DEUTSCHEN LEHNWÖRTER IM TESCHENER DIALEKT DES POLNISCHEN. Thomas Menzel, Gerd Hentschel. 2., ergänzte und korrigierte elektronische Ausgabe. Bundesinstitut für Kultur und Geschichte der Deutschen im östlichen Europa 2005. [Oldenburg: BIS-Verlag 2003 (= Studia Slavica Oldenburgensia 10)] (<http://www.bkge.de/14451.html>) (07.11.2013).
- DLS = DEUTSCHE LEHNWÖRTER IM SLOVENISCHEN. Hildegard Striedter-Temps. Wiesbaden: Harrassowitz 1963.

(b) Sonstige Literatur

- Burch, Thomas/Rapp, Andrea (2007): „Das Wörterbuch-Netz: Verfahren – Methoden – Perspektiven.“ – In: D. Burckhardt, R. Hohls, C. Prinz (Hgg.): *Geschichte im Netz: Praxis, Chancen, Visionen. Beiträge der Tagung .hist 2006*, 607-627. Berlin (=Historisches Forum 10, Teilband I). (http://edoc.hu-berlin.de/histfor/10_I/PHP/Woerterbuecher_2007-10-I.php#007001) (07.11.2013).
- Burnard, Lou/Bauman, Syd (Hgg.) (2007): *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. – Charlottesville, Virginia: TEI Consortium. (<http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html>) (07.11.2013).
- Engelberg, Stefan (2010): „An inverted loanword dictionary of German loanwords in the languages of the South Pacific.“ – In: A. Dykstra, T. Schoonheim (Hgg.): *Proceedings of the XIV EURALEX International Congress (Leeuwarden, 6–10 July 2010)*, 639-647. Ljouwert (Leeuwarden): Fryske Akademy.
- Meyer, Peter/Engelberg, Stefan (2011): „Ein umgekehrtes Lehnwörterbuch als Internetportal und elektronische Ressource: Lexikographische und technische Grundlagen.“ – In: H. Hedeland, Th. Schmidt, K. Wörner (Hgg.): *Multilingual Resources and Multilingual Applications*, 169-174. Hamburg: Universität Hamburg (= Arbeiten zur Mehrsprachigkeit/Working Papers in Multilingualism, Folge B, Nr. 96).
- Meyer, Peter (2013a): „Ein Internetportal für deutsche Lehnwörter in slavischen Sprachen. Zugriffsstrukturen und Datenrepräsentation.“ – In: S. Kempgen, N. Franz, M. Jakiša, M. Wingender (Hgg.): *Deutsche Beiträge zum 15. Internationalen Slavistenkongress, Minsk 2013*, 233-242. München: Otto Sagner (=Die Welt der Slaven. Sammelbände, Band 50).
- Meyer, Peter (2013b): „Advanced graph-based searches in an Internet dictionary portal.“ – In: I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets, M. Tuulik (Hgg.): *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia*, 488-502. Ljubljana, Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut (http://eki.ee/elex2013/proceedings/eLex2013_34_Meyer.pdf) (07.11.2013).
- van der Sijs, Nicoline (2010): *Nederlandse woorden wereldwijd*. – Den Haag: SDU Uitgever.
- Webber, Jim/Eifrem, Emil/Robison, Ian (2013): *Graph Databases*. – Sebastopol, CA: O'Reilly & Associates.