Erschienen in: Kempgen, Sebastian/Wingender, Monika/Franz, Norbert/Jakiša, Miranda (Hrsg.): Deutsche Beiträge zum 15. Internationalen Slavistenkongress, Minsk 2013. - München/Berlin/ Washington: Otto Sagner, 2013. S. 233-242. (Die Welt der Slaven. Sammelbände, Band 50)

EIN INTERNETPORTAL FÜR DEUTSCHE LEHNWÖRTER IN SLAVISCHEN SPRACHEN
Zugriffsstrukturen und Datenrepräsentation

1. Ein Internetportal für Lehnwörterbücher als umgekehrtes Lehnwörterbuch

Vor über 30 Jahren wurde von Ju.N. Karaulov (1979) die Erstellung eines „umgekehrten Lehnwörterbuchs" (*обратный словарь заимствований*) vorgeschlagen, das Entlehnungen aus dem Russischen in zahlreiche Nehmersprachen aus der Sicht der Gebersprache dokumentiert und dementsprechend nach russischen Etyma lemmatisiert ist. Karaulovs Idee wurde für das Russische nicht verwirklicht; eine vergleichbare Idee wurde jedoch unabhängig davon in jüngerer Zeit im Kontext deutschsprachiger Entlehnungen in den südpazifischen Sprachen artikuliert (Engelberg 2010) und ist Gegenstand eines laufenden Forschungsprojektes am Institut für Deutsche Sprache (IDS) in Mannheim.

Konsequent umgesetzt wurde die Idee eines umgekehrten Lehnwörterbuchs bislang nur für das Niederländische: In der Monographie mit dem programmatischen Titel *Nederlandse woorden wereldwijd* (van der Sijs 2010) folgt auf einen ausführlichen einleitenden Text ein sehr umfangreicher Wörterbuchteil (S. 163-723), dessen nach niederländischen Wörtern lemmatisierte Artikel im wesentlichen Wortgleichungen sind; hierbei wird für das niederländische Quellwort immer, für die im Artikel u.U. in mehreren Varianten genannten Lehnwörter – außer der Sprachenzuordnung und ggf. dem Vermerk, dass das Lehnwort nicht mehr gebräuchlich ist – meistens eine kurze Bedeutungsangabe geliefert. Darüber hinaus finden sich noch, wo erforderlich, Informationen zum Entlehnungsweg, falls das betreffende niederländische Wort durch Vermittlung einer anderen Sprache entlehnt wurde. Die Autorin weist selber nachdrücklich auf die – unvermeidliche – Unvollständigkeit und den Pioniercharakter dieser Arbeit hin. Ein erheblicher Teil der Einträge ist anderen lexikographischen Arbeiten entnommen; schon aus Platzgründen findet man jedoch keinerlei Hinweise auf die jeweilige Informationsquelle. In der vorliegenden Printfassung gehen sämtliche weiteren lexikographischen Informationen der herangezogenen Wörterbücher verloren und müssen vom interessierten Benutzer fallweise selber ermittelt werden. In einer Onlinefassung des Wörterbuchs wäre es möglich, solche Hinweise in die Artikel zu integrieren bzw., falls möglich, direkt darauf zu verlinken. Selbst dann wäre es jedoch nicht ohne weiteres möglich, auch z.B. grammatische Informationen oder Angaben zum Entlehnungszeitraum in Suchvorgänge einzubeziehen, solange solche Angaben, die in verschiedenen Lehnwörterbüchern in unterschiedlichen Formaten vorliegen, nicht in vereinheitlichter Form zum Bestandteil der lexikographischen Datenbasis des umgekehrten Lehnwörterbuchs gemacht werden.

Angesichts dieser Schwierigkeiten wird am Institut für Deutsche Sprache eine alternative Strategie der Erstellung eines umgekehrten Lehnwörterbuchs zu deutschen lexikalischen Entlehnungen in anderen Sprachen verfolgt: In einem Internetportal werden bereits vorhandene Wörterbücher zu deutschen Lehnwörtern in anderen Sprachen in digital aufbereiteter, inhaltlich unveränderter Form als Einzelwerke zugänglich gemacht und zusätzlich lexikographisch so miteinander vernetzt, dass in einem separaten umgekehrten Lehnwörterbuch (Wörterbuch der

deutschen Herkunftswörter des Portals) ausgehend von deutschen Etyma nach zugehörigen Entlehnungen in den Nehmersprachen aller integrierten Wörterbücher gesucht werden kann. Durch diese Herangehensweise geht für komplexere Suchanfragen keine lexikographische Information verloren, der Benutzer hat bei Suchen im umgekehrten Lehnwörterbuch stets direkten Zugriff auf die Quelle der betreffenden Informationen.

Im Rahmen eines Pilotprojektes[1] wurde die grundsätzliche Portalarchitektur konzipiert und implementiert, und danach wurden zunächst drei einzelsprachige Lehnwörterbücher in das Portal integriert. Das umfangreichste davon ist mit 2447 Artikeln das bereits vorher online veröffentlichte *Wörterbuch der deutschen Lehnwörter in der polnischen Schrift- und Standardsprache* (de Vincenz & Hentschel 2010), im folgenden *WDLT* abgekürzt. Ebenfalls online ist bereits seit längerem das *Wörterbuch der deutschen Lehnwörter im Teschener Dialekt des Polnischen* (Menzel & Hentschel 2005) mit 839 Artikeln, im folgenden kurz *WDLT*, verfügbar. Schließlich ist der Wörterbuchteil der Monographie *Deutsche Lehnwörter im Slovenischen* (Striedter-Temps 1963), im weiteren kurz *ST*, integriert worden, der ohne reine Verweiseinträge 1568 Artikel umfasst.

Datenbankseitig wird jeder Artikel eines jeden Lehnwörterbuchs durch ein separates digitales Dokument in einer wörterbuchspezifischen Anwendung des XML-Auszeichnungsformates repräsentiert. Im Falle der beiden polnischen Wörterbücher wurden die Artikel im Rahmen einer Kooperation mit dem Institut für Slavistik der Universität Oldenburg (Lehrstuhl für slavistische Sprachwissenschaft, Prof. Dr. Gerd Hentschel) bereits im XML-Format zur Verfügung gestellt. Für ST wurde ein kombiniertes Verfahren gewählt: Zum einen werden die Einzelartikel als gescannte Bilder im Online-Wörterbuch zur Verfügung gestellt, zum anderen wurden zentrale lexikographische Informationen der Einzelartikel in aufwändiger manueller Arbeit in vorgefertigte Textverarbeitungsformulare extrahiert und dann automatisiert in XML-Dokumente umgewandelt.

Das Resultat des Pilotprojektes ist das „Lehnwortportal Deutsch", das unter der Adresse <http://lwp.ids-mannheim.de> seit November 2012 frei online zugänglich ist. Die für das Lehnwörterbuchportal am IDS neu entwickelte Webapplikationssoftware ist modular konzipiert, so dass leicht weitere lexikographische Ressourcen integriert werden können. Das IDS ist insbesondere bestrebt, den Kreis der betrachteten slavischen Nehmersprachen zu vergrößern. Da am IDS keine Möglichkeit zur Erstellung neuer Wörterbücher zu deutschen Lehnwörtern in anderen Sprachen besteht, ist das Institut darauf angewiesen, z.B. im Rahmen von Kooperationen fundiertes lexikographisches Material zur Verfügung gestellt zu bekommen oder aber aktuell laufende lexikographische Projekte zu begleiten. Ein Projekt zu deutschem lexikalischen Material, das durch polnische Vermittlung in die ostslavischen Sprachen entlehnt worden ist, ist zur Zeit in Vorbereitung. Es ist überdies beabsichtigt, zu einem späteren Zeitpunkt auch Entlehnungen in nichtslavische Sprachen zu berücksichtigen.

Das Portal bietet unter der URL <http://lwp.ids-mannheim.de/doc/portal/start> ausführliche Benutzungshinweise an. Die folgenden Ausführungen wiederholen diese Erläuterungen nicht und gehen auch nicht auf die wissenschaftlichen Nut-

[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.

zungsmöglichkeiten des Portals ein (vgl. dazu Meyer & Engelberg 2011), sondern diskutieren vertiefend die Frage, welche *Zugriffsstrukturen* ein Lehnwörterbuchportal und speziell ein umgekehrtes Lehnwörterbuch sinnvollerweise anbieten sollte und welche lexikographischen sowie technischen Voraussetzungen dafür vorliegen müssen. Dazu werden in Abschnitt 2 zunächst Anforderungen an Zugriffsstrukturen formuliert, und daraus wird eine geeignete Organisation der lexikographischen Datenbasis abgeleitet, nämlich als ein wörterbuchübergreifendes Netzwerk von Relationen zwischen Etymon- und Lehnwortformen. Abschnitt 3 erläutert die lexikographisch-technische Erstellung dieses Netzwerkes, Abschnitt 4 abschließend die darauf basierende konkrete Realisierung der Zugriffsstrukturen im „Lehnwortportal Deutsch".

2. Anforderungen an Zugriffsstrukturen in einem Portal für Lehnwörterbücher

Bereits Karaulov (1979: 560-562) diskutiert verschiedene Zugriffsweisen (входы) auf ein umgekehrtes Lehnwörterbuch; neben dem traditionellen makrostrukturellen Zugang über eine alphabetisch geordnete Liste der Lemmata (Etyma) werden Zugriffsmöglichkeiten über thematische (wortfeldbezogene, onomasiologische) Einordnung sowie über Entlehnungshäufigkeit (Zahl der Sprachen, in die ein Lexem entlehnt wurde), Zielsprache, Zeitraum der Entlehnung, diasystematischen Status (z.B. Register) und Entlehnungstyp (Lehnwort i.e.S. versus Lehnübersetzung, -übertragung, ...) genannt.[2]

Bereits die Erstellung einer Lemmaliste für ein umgekehrtes Lehnwörterbuch auf der Grundlage von herkömmlichen, nehmersprachlich lemmatisierten Lehnwörterbüchern ist jedoch mit komplexen lexikographischen und philologischen Problemen behaftet, die Karaulov nicht anspricht. So wird beispielsweise im WDLP s.v. *lichtarz* als Etymon eine mhd. Form *liuhtaere* angegeben; die nhd. Fortsetzung *Leuchter* taucht nur (gewissermaßen zufällig) in der Bedeutungsangabe des Etymons auf. In ST wird dagegen als Etymon für *lajhter* nhd. *Leuchter* und mhd. *liuhtære* (mit gegenüber WDLP abweichender Ligaturschreibung des Digraphen *ae*) angegeben, wobei das Lexem gemäß einer nicht dokumentierten Konvention von ST i.a. in der erstgenannten, hier: neuhochdeutschen, Form entlehnt wurde. Schon an diesem sprachhistorisch sehr einfachen Fall wird deutlich, dass eine mechanische Lemmatisierung nach den in den einzelnen Lehnwörterbüchern genannten Etymonformen nicht zielführend ist. Interessierte Laien sollten über ein 'normalisiertes' Lemma *Leuchter* im umgekehrten Lehnwörterbuch auch direkt auf polnisch *lichtarz* geführt werden; Experten sollten in einem Suchformular jedoch auch direkt nach einer mittelhochdeutschen Form wie *liuhtære* suchen können. Es ist also sinnvoll, zwischen den *normalisierten* Lemmata des umgekehrten Lehnwörterbuchs (im folgenden als *Metalemmata* bezeichnet) und den in den Artikeln der einzelnen Lehnwörterbücher konkret genannten Etymonformen zu unterscheiden. 'Normalisierung' bedeutet im Falle des Lehnwortportals Deutsch die Wahl einer diasystematischen Variante, die – und dies ist hier be-

[2] Schon Karaulov schlägt angesichts der Vielzahl wünschenswerter Zugriffsweisen vor, die Beschränkungen eines gedruckten Wörterbuchs durch Verwendung einer Datenbank zu überwinden: „Однако книга — это не единственный путь построения словаря, он может существовать и в виде лексикографического банка данных" (1979: 562). Nur wenig später wurde mit dem von vornherein für eine relationale Datenbank konzipierten WDLP Pionierarbeit in genau dieser Richtung geleistet (Hentschel 1989).

wusst vage formuliert – möglichst wenig vom Ideal eines heute gebräuchlichen neuhochdeutschen standardsprachlichen Lexems abweicht. Im gegebenen Beispiel ist offenbar *Leuchter* ein geeignetes Metalemma, dem im WDLP die Etymonform *liuhtaere* und in ST die Etymonform *Leuchter* zuzuordnen ist. Häufig ist aber kein klarer Kandidat für die Normalisierung in Sicht, etwa wenn das Etymon aus einer älteren Sprachstufe stammt und keine neuhochdeutsche Fortsetzung bzw. Entsprechung hat, wie mhd. *lît* 'Obstwein' für sloven. *lit.*[3] Ähnlich problematisch sind Fälle, in denen die naheliegende nhd. Entsprechung eines genannten Etymons morphologisch mehr oder weniger abweicht, beispielsweise durch andere Suffigierung. Ganz allgemein ist die Zuordnung von in verschiedenen Artikeln und/oder Wörterbüchern genannten Etymonformen, die unterschiedlichen Sprachstufen, Dialekten usw. angehören, zu einem und demselben Metalemma eine eigenständige lexikographische Leistung, die in jedem Einzelfall sprachhistorisch zu begründen ist. Durch die Zuordnung zu einem Metalemma werden die Etyma etymologisch miteinander identifiziert, d.h. als diasystematische Varianten voneinander ausgewiesen.

Es ist wichtig festzuhalten, dass die Metalemmata des umgekehrten Lehnwörterbuchs nicht einfach jeweils bestimmten Artikeln als ganzes zugeordnet werden können, sondern den zugehörigen konkreten Etyma in diesen Artikeln. So werden in ST häufig in einem Artikel mehrere slovenische Lehnwörter mit ihren zugehörigen deutschen Etyma behandelt: Der Artikel *antlati* etwa bespricht neben dem Verb *antlati* mit seinen Etyma mhd. *handeln*, ahd. *hantalôn* (anzusetzendes Metalemma: *handeln*) auch das Substantiv *(h)antel*, dessen Etyma ahd. **hantal/*hantel*, mhd. **handel* zu einem Metalemma *Handel* zu stellen sind. Für komplexere Suchen im Portaldatenbestand wird man fordern müssen, dass die korrekte Zuordnung der einzelnen Etyma zu den zugehörigen Lehnwörtern (sowie beispielsweise auch die Zuordnung von Varianten und Derivaten zum jeweiligen Lehnwort) für solche Artikel erhalten bleibt. Es ist für erweiterte Suchfunktionalitäten außerdem wünschenswert, dass die Derivationsbeziehung zwischen den Metalemmata *Handel* und *handeln* ebenfalls datenbankseitig abgebildet wird, da ansonsten nur eine wenig aussagekräftige zeichenbasierte Suche möglich wäre. Die Metalemmaliste muss daher sinnvollerweise in manueller lexikographischer Arbeit um Relationen (Vernetzungen) zwischen ihren einzelnen Elementen angereichert werden. Insgesamt ergibt sich aus dem Gesagten die Forderung nach einer *relationalen, wortformbezogenen Repräsentation* der lexikographischen Daten des Portals: Für Suchprozesse müssen diese Daten, unabhängig von den Idiosynkrasien der Makro- und Mikrostrukturen der einzelnen Wörterbücher, als Menge von Beziehungen zwischen Wortformen, einschließlich der Metalemmata, zur Verfügung gestellt werden. Hierbei bezieht sich der im weiteren durchgehend verwendete neutrale Terminus 'Wortform' auf die konkreten, in Wörterbüchern gebuchten Formen von Metalemmata, Etyma und Lehnwörtern, die im Allgemeinen nicht Lemmata der betreffenden Artikel sind und auch nicht in allen Fällen (z.B. nicht bei orthographischen Varianten) Zitierformen von jeweils unterschiedlichen Lexemen sind.

Die weiteren von Karaulov genannten Zugriffsmöglichkeiten können in einer Webanwendung durch eine *wörterbuchübergreifende* Suchfunktionalität abgebil-

[3] Die Sache wird im vorliegenden Fall daduch kompliziert, dass manche Wörterbücher des Nhd. immerhin ein dialektales und veraltetes *Leitgeb(er)* 'Wirt' kennen.

det werden, bei der Boolesche Verknüpfungen von Suchkriterien in ein Formular eingegeben werden. Sowohl inhaltlich als auch technisch setzen wörterbuchübergreifende Suchen jedoch eine *einheitliche Datenrepräsentation* für die Angaben aller Lehnwörterbücher des Portals voraus. Schon die drei bislang integrierten Wörterbücher unterscheiden sich jedoch beträchtlich hinsichtlich der Mikrostruktur, der Angabetypen und der Angabeformate ihrer Artikel: das WDLP verfügt über eine hoch elaborierte und formalisierte Mikrostruktur und legt einen Schwerpunkt auf Belege und deren Datierungen. Die Mikrostruktur von WDLT-Artikeln ist deutlich weniger komplex, mit einem Schwerpunkt auf lexikalischen Parallelen der besprochenen Lehnwörter in benachbarten Slavinen und auch im Deutschen; Angaben zu Wortart, Genus bei Substantiven usw. fehlen. Die sehr informell strukturierten Artikel von ST beleuchten vor allem Fragen der historischen Phonologie und liefern in sehr variierendem Umfang zusätzliche Informationen zur Grammatik, zu weiteren Sprachstufen des Etymons usw.

Die verschiedenen Mikrostrukturen der bislang eingebundenen Wörterbücher spiegeln sich in dem sehr unterschiedlichen strukturellen Aufbau der XML-Artikelinstanzen, dem sogenannten XML-Schema der drei Wörterbücher, wider. Natürlich wäre es prinzipiell denkbar, ein sehr granulares und allgemeines XML-Schema zu entwickeln, auf das die Mikrostruktur eines jeden einzubindenden Wörterbuchs abzubilden wäre, und so die gewünschte einheitliche Datenrepräsentation zu erzeugen. Die enorme Vielfalt denkbarer Mikrostrukturen in lehnwortlexikographischen Ressourcen macht dies jedoch zu einem fast unmöglichen Unterfangen, wenn der Anspruch besteht, dass bei der Überführung in dieses portalweite Datenformat keine ressourcenspezifischen Angaben verloren gehen. In der Praxis liefe dies darauf hinaus, das portalweite XML-Schema z.B. bei jedem Auftreten eines nicht antizipierten Angabetyps eines neu zu integrierenden Lehnwörterbuch neu zu definieren und alle bereits vorhandenen XML-Instanzen entsprechend anzupassen.

Hinzu kommt die Schwierigkeit, dass für portalweite, wörterbuchübergreifende Suchvorgänge die sehr heterogenen und unterschiedlich granularen Angaben der Ausgangswörterbücher zu Sprachstufen bzw. Dialekten, aus denen bzw. in die entlehnt wurde, sowie zum Entlehnungszeitraum und zu grammatischen Eigenschaften von Etyma und Lehnwörtern in ein einheitliches konzeptuelles Schema überführt werden müssen. Da es möglich sein muss, Artikel einzelner Wörterbücher des Portals zu aktualisieren und da Benutzer die Originalwerke in unveränderter Form konsultieren können sollen, muss die für portalweite Suchen hinzugefügte Information getrennt von den Artikeldaten verwaltet werden.

Die im Lehnwortportal Deutsch gewählte Lösung für die skizzierte Problemlage besteht darin, zum einen mit wörterbuchspezifischen XML-Schemata zu arbeiten und die XML-Instanzen der Einzelwörterbücher nicht mit portalrelevanter Information anzureichern;[4] zum anderen wird für die Zwecke einer einheitlichen

[4] Es wird jedoch, um weitgehend automatisierte Verarbeitung zu ermöglichen, vom XML-Schema für die Einzelartikel eines jeden Wörterbuchs jeweils verlangt, dass es möglichst weitgehend von Layout- und Präsentationsaspekten abstrahiert, etwa im Sinne des *lexical view* der TEI.dictionaries-Richtlinien (Burnard & Bauman 2007). Es ist jedoch wesentlich festzuhalten, dass die Anwendung einer TEI-konformen Kodierung allein nicht die gewünschte Einheitlichkeit der lexikographischen Datenrepräsentation gewährleistet.

Datenrepräsentation eine weitere Datenhaltungsschicht mit portalweit einheitlichen Angabeformaten definiert, die insbesondere für wörterbuchübergreifende Suchanfragen verwendet wird und außerdem die oben formulierte Forderung nach einer relationalen, wortformbezogenen Datenrepräsentation erfüllt.

Die zusätzliche Datenhaltungsschicht muss nach dem vorher Gesagten ein 'Netzwerk' von Beziehungen zwischen Wortformen sein, wobei jeder Wortform z.B. diasystematische und grammatische Informationen in einem einheitlichen Angabeformat zugeordnet sind. Zur Zeit werden im Portal folgende Beziehungen unterschieden:

- Etymon/Lehnwort *x* ist (orthographische, phonologische, ...) *Variante* von Etymon/Lehnwort *y*.
- Metalemma/Etymon/Lehnwort *x* ist morphologische *Ableitung* bzw. *Kompositum* zu Metalemma/Etymon/Lehnwort *y*.
- Wortform *x* ist Etymon *y* als *Lehnwort* zugeordnet.
- Wortform *x* ist Metalemma *y* als (wörterbuchartikelspezifisches) *Etymon* zugeordnet.
- Wortform *x* ist *lexikalische Parallele* zu Lehnwort *y* (vgl. die obigen Erläuterungen zum WDLT).

Formal lässt sich das Netzwerk mit den Mitteln der Graphentheorie als *gerichteter azyklischer Graph* (azyklischer Digraph) beschreiben (Bang-Jensen & Gutin 2009: 32-34), dessen Knoten einzelne Wortformen repräsentieren. Die genannten Beziehungstypen zwischen je zwei Wortformen sind mathematisch als antisymmetrische und irreflexive Relationen modellierbar und werden durch gerichtete Kanten zwischen den entsprechenden Knoten repräsentiert. Dabei darf es schon aus inhaltlichen Gründen keine Schleifen (Zyklen) geben, wenn man in diesem Graphen Pfade verfolgt; durch die Azyklizität wird außerdem die Semantik und technische Realisierung von Suchvorgängen im Graphen deutlich vereinfacht. In der Aufzählung oben kann die Richtung wahlweise immer von *x* nach *y* oder aber immer von *y* nach *x* gedacht werden; für die weiteren Ausführungen sei die letztere Betrachtungsrichtung gewählt, bei der die Kanten im Graphen von Metalemmata zu Etyma, von dort weiter zu Lehnwörtern und schließlich zu deren Varianten, Derivaten usw. führen. Die graphenbasierte Modellierung wird es in künftigen Ausbaustufen erlauben, Entlehnungsketten in der Datenbasis des Portals abzubilden, die über beliebig viele Sprachen laufen können. Dies wird bereits für die erwähnte erste Erweiterung des Portals, die polnisch vermittelte deutsche Entlehnungen im Ostslavischen betrifft, von entscheidender Bedeutung sein.

Eine graphentheoretische Repräsentation auf der Ebene von Einzelwortformen wird in der hier beschriebenen Form in der Internetlexikographie ansonsten bislang noch nicht eingesetzt; das Wörterbuchnetz < http://woerterbuchnetz.de/ > des Trierer Kompetenzzentrums für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften verwendet zwar einen azyklischen Digraphen, die Knoten des Graphen sind jedoch ganze Wörterbuchartikel (Burch & Rapp 2007).

3. Erstellung des gerichteten azyklischen Graphen: lexikographische und technische Aspekte

Der gerichtete azyklische Wortformengraph des Portals muss aus den lexikographischen Informationsquellen des Portals durch einen weitgehend automatisierten Prozess gewonnen werden. Es gibt zwei Arten von Informationsquellen:

Zum einen wird jeder Artikel eines jeden Lehnwörterbuchs durch ein XML-Dokument repräsentiert, aus dem jeweils ein entsprechender Teilgraph automatisiert extrahiert werden kann, welcher alle im Artikel genannten Etyma und Lehnwörter, ggf. mit deren Varianten, Derivaten usw., und ihre Relationen zueinander umfasst. Für diesen Extraktionsvorgang muss auf der Grundlage der wörterbuchspezifischen XML-Artikelstrukturen für jedes Lehnwörterbuch jeweils ein Algorithmus formuliert und implementiert werden, der bestimmte Fragmente eines XML-Dokuments jeweils in bestimmte Relationen zwischen Wortformen 'übersetzt'. Voraussetzung dafür ist eine hinreichend granulare, an inhaltlichen Strukturen orientierte XML-Repräsentation für alle Wörterbücher (vgl. Fußnote 4).

Zum anderen müssen die Metalemmata mitsamt ihren Verknüpfungen mit den Etyma sowie untereinander in den Graphen integriert werden. In der derzeit für das Lehnwortportal gewählten Lösung wird die Metalemmaliste mit Hilfe einer zu diesem Zweck entwickelten Software erstellt, bearbeitet und verwaltet sowie als XML-Dokument gespeichert, aus dem wiederum die entsprechenden Teilgraphen extrahiert werden können. Das Programm stellt dem Lexikographen eine aus den XML-Artikeldokumenten gewonnene wörterbuchübergreifende Liste von Etyma sowie eine Liste aller bislang angelegten Metalemmata zur Verfügung und ermöglicht das Erstellen, Kontrollieren und Löschen von Beziehungen zwischen den Elementen der beiden Listen per Mausklick. Schon bei wenigen tausend Etyma wird es unumgänglich, schnell in beiden Listen suchen zu können. Die Software beherrscht mit Einschränkungen auch „mediävistisch informierte" unscharfe Suchen, die z.B. niederdeutsche oder mittelhochdeutsche Entsprechungen eines eingegebenen neuhochdeutschen Wortbestandteils finden. Ferner können Metalemmata miteinander vernetzt oder nachträglich zusammengelegt und somit etymologisch gleichgesetzt werden.

Im Ergebnis enthält die Datenbasis eine große Zahl von Teilgraphen, deren Pfade jeweils mit einem bestimmten Metalemma (z.B. *Leuchter*) beginnen. Jedem Metalemma sind Etyma aus verschiedenen Artikeln des Portals zugeordnet (*Leuchter*, *liuhtære*, *liuhtaere*, ...), diesen wiederum die zugehörigen Lehnwörter usw. Die genannten Teilgraphen sind miteinander im allgemeinen nicht verbunden, es sei denn, es bestehen Beziehungen zwischen Metalemmata. Für jedes Metalemma kann sich der interessierte Benutzer den zugehörigen Teilgraphen im Portal als interaktive Graphik anzeigen lassen.

Bei jeder Änderung an den zugrundeliegenden Datenquellen muss der Gesamtgraph in der Datenbank neu aufgebaut werden. Es wäre auch denkbar, den Graphen als eigenständigen Bestandteil der lexikographischen Datenbasis zu behandeln und unabhängig von den ursprünglichen Quellen nach Bedarf zu erweitern und zu modifizieren. Ein solches Vorgehen würde jedoch Konsistenzprüfungen und das Nachvollziehen der Revisionsgeschichte des Graphen sehr erschweren.

Wie bereits betont, entspricht jeder Knoten des Graphen einer konkreten Wortform, zu der lexikographische Informationen in einem wörterbuchübergreifend einheitlichen Format abgespeichert sind. Die einer bestimmten Wortform zugehörigen Informationen kann sich der interessierte Benutzer konkret in der erwähnten

interaktiven Graphik für das zugehörige Metalemma oder aber auch im entsprechenden Lehnwörterbuchartikel anzeigen lassen. In der derzeitigen Ausbaustufe des Portals werden beim automatisierten Extrahieren des Graphen durch einen automatischen Prozess jeder Wortform vereinheitlichte diasystematische und morphosyntaktische Informationen zugeordnet; in Zukunft könnten z.B. noch phonologische oder morphologische Analysen hinzukommen. Erst so wird es möglich, beispielsweise bei einer Suche nach Artikeln mit hochdeutschen oder oberdeutschen Etyma auch solche Etyma zu berücksichtigen, die im Originalwörterbuch spezifischer als „bairisch" ausgewiesen sind. Zu jeder deutschen Wortform wird auch eine normalisierte, vereinfachte graphematische Repräsentation gespeichert; auf diese Weise werden in den erweiterten Suchen auch vereinfachte Schreibweisen ohne Diakritika möglich – bei der Suche nach einem Etymon *drot* wird auch die Form *drǭt* gefunden.

Der gerichtete Graph wird derzeit in einer relationalen Datenbank verwaltet; die Knoten, also die Einzelwortformen mitsamt ihren Eigenschaften, werden jeweils als mit einer ID versehene Datensätze (Zeilen) einer Tabelle gespeichert, die Kanten (Relationen) in einer zweiten Tabelle als geordnete und etikettierte Paare von Knoten-IDs. Aus Gründen der höheren Performanz bei Suchen werden dabei auch alle indirekten Beziehungen zwischen Knoten – der sogenannte transitive Abschluss des Graphen – in der Relationentabelle abgelegt.[5]

4. Realisierung der Zugriffsstrukturen im „Lehnwortportal Deutsch"

Die im Portal eingestellten Lehnwörterbücher verfügen über eine einheitliche Benutzerführung; an herkömmlichen, lemmabasierten Zugriffsmöglichkeiten in der gewöhnlichen Artikelansicht wird neben einer wörterbuchspezifischen Alphabetleiste und einem scrollbaren Ausschnitt aus der jeweiligen Lemmaliste noch eine „Autocomplete"-Funktion zum schnellen Eingeben von Lemmata angeboten. Diese Zugriffsstrukturen verwenden nicht den gerichteten Graphen, der ja von der Artikelstruktur der Wörterbücher abstrahiert, sondern eine separate Tabelle der Lemmata aller Wörterbücher (samt zugehörigen XML-Dokumenten) sowie der Metalemmata. Die Artikelanzeige selbst wird von der Webanwendung auf gängige Weise dynamisch erzeugt, indem das zugehörige XML-Dokument auf wörterbuchspezifische Weise mittels einer sogenannten XSL-Transformation in ein HTML-Fragment umgewandelt wird (vgl. Müller-Spitzer 2007: 212ff.).

Insbesondere an Laien und Gelegenheitsbesucher des Portals richtet sich die auf allen Portalseiten zur Verfügung stehende Schnellsuche nach (beliebigen Bestandteilen von) normalisierten deutschen Herkunftswörtern. Der Nutzer erhält auf der Ergebnisseite sämtliche Metalemmata, die (modulo Groß-/Kleinschreibung) mit der eingegebenen Zeichenfolge beginnen, enden, sie enthalten bzw. mit ihr identisch sind. Auf der Ergebnisseite gelangt der Benutzer von jedem genannten Metalemma per Hyperlink zum zugehörigen Artikel des umgekehrten Lehnwörterbuchs. Die Artikel des umgekehrten Lehnwörterbuchs werden dynamisch aus den Informationen des gerichteten Graphen erzeugt; sie zeigen im wesentli-

[5] Auf Details der technischen Implementierung wird hier nur am Rande eingegangen; ausführlicher dazu Meyer & Engelberg (2011). Alternativ können gerichtete Graphen auch in dedizierten Graphendatenbanken wie *Neo4j* <http://www.neo4j.org/> verwaltet werden.

chen, nach Lehnwörterbuch und (verlinktem) Artikel geordnet, die dem Metalemma zugeordneten Etyma.

Umgekehrt findet sich am Ende eines jeden Lehnwörterbucheintrags ein Verweis auf die zugehörigen Einträge des Herkunftswörterbuchs samt der Möglichkeit, sich die nach zugeordnetem Metalemma geordneten Wortformen des Eintrags samt der dazu im Graphen verfügbaren Angaben anzeigen zu lassen. Die Lehnwörterbücher und das Herkunftswörterbuch haben also eine wechselseitige Verweisstruktur, die auch Laien den Umgang mit dem Portal erleichtert.

Vor allem an Fachleute sind die erweiterten Suchmöglichkeiten des Portals gerichtet. In der portalweiten Suche[6] unter der URL <http://lwp.ids-mannheim.de/search/meta> kann der Nutzer sowohl für Etyma (wahlweise mit Einschluss der normalisierten Formen der entsprechenden Metalemmata) als auch für etymologisch zugehörige Lehnwortformen Bestandteile ihrer graphematischen Form, Bestandteile der zugehörigen Bedeutungserläuterungen, diasystematische Einordnung und grammatische Informationen spezifizieren und mit Booleschen Operatoren kombinieren. Für jedes Kriterium kann man zusätzlich einen Skopus angeben; so ist es möglich, Eigenschaften von anderen, etymologisch zugehörigen Wortformen im selben Artikel oder aber in beliebigen Wörterbüchern des Portals zu spezifizieren. Auf diese Weise werden äußerst komplexe Suchanfragen möglich. Ein Beispiel ist die Anfrage „Suche Artikel, in denen ein mittelhochdeutsches feminines Etymon auf *-unge*, dessen Bedeutungsangabe *-geld-* oder *-handel-* enthält, zu einem polnischen maskulinen Lehnwort auf *-unek* gehört, zu dem im selben Artikel außerdem ein etymologisch zugehöriges Wort auf *-owy* genannt wird und für das es ein etymologisch verwandtes Wort im Slovenischen gibt". Diese Anfrage liefert den Artikel *szacunek* im WDLT mit den passenden Etymon-Lehnwortpaaren *schatzunge* → *oszacunek* und *schatzunge* → *szacunek*.[7]

Die erweiterte Suche operiert ausschließlich auf dem gerichteten Wortformengraphen; das oben mehrfach erwähnte zentrale Kriterium der *etymologischen Zu(sammen)gehörigkeit* zweier Wortformen wird hierbei durch die Eigenschaft dieser Wortformen operationalisiert, im Graphen mit einem gemeinsamen Metalemma verbunden zu sein. – In zukünftigen Ausbaustufen des Portals ist auch die Verwendung einer deklarativen Abfragesprache denkbar, die direkt die Suche nach beliebig komplexen Teilgraphkonfigurationen ermöglicht.

Zusammenfassend lässt sich festhalten, dass die Vielfalt und Komplexität der Suchmöglichkeiten des Portals schon bei den wörterbuchspezifischen Suchen für die beiden bereits vorher online veröffentlichten Wörterbücher zum Polnischen weit über die bisher verfügbaren Zugriffs- und Recherchemöglichkeiten hinausgeht; erst durch die innovative graphenbasierte, wörterbuchübergreifende Vernetzung auf Wortformenebene wird jedoch aus dem Portal mehr als nur eine Zusammenstellung von heterogenen lexikographischen Einzelressourcen, in denen man wörterbuchübergreifend allenfalls zeichenbasierte Suchen durchführen könnte.

---

[6] Für die Einzelwörterbücher gibt es separate erweiterte Suchen mit zusätzlichen, wörterbuchspezifischen Suchoptionen; so wird für das WDLP eine Suche nach Erstbelegsdatum angeboten, für das WDLT eine Suche nach lexikalischen Parallelformen aus anderen polnischen Varietäten, dem Tschechischen sowie deutschen Dialekten.

[7] Im Artikel wird das Adjektivderivat *szacunkowy* genannt; zugehörig ist das slovenische *šacinga* im ST-Artikel *šacati*.

Literaturverzeichnis

Bang-Jensen, J., Gutin, G. Z. ²2009. *Digraphs: theory, algorithms and applications.* London.

Burch, T., Rapp, A. 2007. Das Wörterbuch-Netz: Verfahren – Methoden – Perspektiven. In: Burckhardt, D., R. Hohls, C. Prinz (Hrsg.): *Geschichte im Netz: Praxis, Chancen, Visionen. Beiträge der Tagung .hist 2006* (= Historisches Forum 10/I). Berlin, 607-627. Online: <http://edoc.hu-berlin.de/histfor/10_I/PHP/Woerterbuecher_2007-10-I.php#007001>.

Burnard, L., Bauman, S. 2007. (Eds.) *TEI P5: Guidelines for Electronic Text Encoding and Interchange.* Charlottesville/Virginia. Online: <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html>.

Engelberg, S. 2010. An inverted loanword dictionary of German loanwords in the languages of the South Pacific. In: Dykstra, A., T. Schoonheim (eds.): *Proceedings of the XIV EURALEX International Congress (Leeuwarden, 6–10 July 2010).* Ljouwert (Leeuwarden), 639-647.

de Vincenz, A., Hentschel, G. 2010. *Wörterbuch der deutschen Lehnwörter in der polnischen Schrift- und Standardsprache. Von den Anfängen des polnischen Schrifttums bis in die Mitte des 20. Jahrhunderts* (= Studia Slavica Oldenburgensia 20). Oldenburg. Online: <http://diglib.bis.uni-oldenburg.de/bis-verlag/wdlp/>.

Hentschel, G. 1989. A dictionary of language contact as a relational data base. In: McCrank, L. J. (ed.): *Data bases in the humanities and social sciences. 4. Proceedings of the ICDBHSS, Montgomery, Al., July 1987.* Medford, 293-302.

Karaulov, J.N. 1979. Obratnyj slovar' zaimstvovanij kak sposob isučenija lingvoėkologii. *Izvestija Akademii Nauk SSSR. Serija Literatury i Jazyka* 38/6, 552-562.

Menzel, T., Hentschel, G. ²2005. *Wörterbuch der deutschen Lehnwörter im Teschener Dialekt des Polnischen.* 2., ergänzte und korrigierte elektronische Ausgabe: <http://www.bkge.de/14451.html>.

Meyer, P., Engelberg, S. 2011. Ein umgekehrtes Lehnwörterbuch als Internetportal und elektronische Ressource: Lexikographische und technische Grundlagen. In: Hedeland, H., T. Schmidt, K. Wörner (eds.): *Multilingual Resources and Multilingual Applications. Proceedings of the Conference of the German Society for Computational Linguistics and Language Technology (GSCL) 2011* (= Arbeiten zur Mehrsprachigkeit/ Working Papers in Multilingualism, Folge B, Nr. 96). Hamburg, 169-174.

Müller-Spitzer, C. 2007. *Der lexikographische Prozess. Konzeption für die Modellierung der Datenbasis* (= Studien zur deutschen Sprache 42). Tübingen.

Striedter-Temps, H. 1963. *Deutsche Lehnwörter im Slovenischen.* Wiesbaden.

van der Sijs, N. 2010. *Nederlandse woorden wereldwijd.* Den Haag.

Institut für Deutsche Sprache, Mannheim — Peter Meyer
(meyer@ids-mannheim.de)