Das Russische Universalwörterbuch (RUW) – didaktisches, translatorisches und linguistisches Hilfsmittel

Bernd Bendixen, Horst Rothe

Universität Leipzig, Philologische Fakultät / Universitätsrechenzentrum
Augustusplatz, 04105 Leipzig, Germany
E-mail: bendixen@rz.uni-leipzig.de, horst.rothe@rz.uni-leipzig.de

Kurzfassung

Das zweisprachige russisch-deutsche Universalwörterbuch versucht, in der Breite des erfassten vorwiegend allgemeinsprachlichen Wortkorpus und in dessen intralingualer wie interlingualer Beschreibung neue Maßstäbe für solche elektronische Wörterbücher zu setzen, die als Nachschlagewerk unterschiedlichsten Nutzerbedürfnissen didaktischer, translatorischer und linguistischer Ausrichtung gerecht werden können. Beschrieben werden die Kriterien einer solchen Anlage und ihre programmtechnische Umsetzung.

1 Problemstellung

Computergestützte Wörterbücher sind längst keine Neuheit mehr, und ihr Einsatz gehört zur alltäglichen Praxis all derer, die Auskunft zu sprachlichen Gegebenheiten oder sprachenpaarbezogenen Entsprechungen suchen. Dabei teilen viele Computerwörterbücher das Schicksal der Printfassungen – nicht nur, dass sie einen oft im Stich lassen, ein Mangel, den sicher jeder kennt, der insbesondere mit zweisprachigen Wörterbüchern arbeitet: das Wort oder zumindest die Wortbedeutung, die für eine lexikalische Einheit im vorgefundenen Kontext zutreffen könnte, ist gerade nicht enthalten, wofür die unterschiedlichsten Gründe ausschlaggebend sein könnten – vielleicht gehört das entsprechende Wort zum sehr speziellen Fachwortschatz, den ich in einem allgemeinsprachlichen Wörterbuch nicht finden kann, oder ich habe gar nicht erkannt, welche Lexikoneinheit sich hinter einer im Text vorgefundenen Zeichenfolge verbirgt. Wenn letzteres auch eher für den Anfänger zutreffen mag, wenn beispielsweise возьму (*ich nehme*) nicht als Wortform des sehr irregulär konjugierten взять (*nehmen*) erkannt wird, verbirgt sich doch hinter beiden Problemen – auch dem hier nicht näher zu illustrierenden Fachwortschatzproblem, das beispielsweise beim medizinischen Terminus "Hüftkopfprothese" augenscheinlich wird – eine Fragestellung, die eigentlich nur den Beschränkungen des Printmediums geschuldet ist und die – überkommen von den Konventionen der Druckfassung – nutzerseitig eine oft stille Übereinkunft voraussetzt, Wortformen auf Wörterbuchformen (Nennformen, Nominativ-Singular-Formen) zurückzuführen und diese auch noch hinsichtlich der Zugehörigkeit zum Allgemein- oder zum Fachwortschatz zu bewerten und damit eine Wahrscheinlichkeitsprognose zu treffen, in welchem der schwer zählbaren und immer weiter ausufernden Einzelwörterbücher denn nun das mich interessierende Wort vorzufinden sein wird.

Vom Prinzip her könnten für ein universelleres elektronisches Wörterbuch sowohl die Beschränkungen auf die wörterbuchformengerechte Stichwortaufnahme als auch auf die Trennung von Allgemein- und von Fachwortschatz entfallen, sofern das Niederreißen dieser überkommenen Barrieren für den Nutzer nicht zu neuen Problemen führt – sicher nicht bei der Wortformenberücksichtigung, eher schon bei der Mischung von Allgemein- und Fachwortschatz, die die Gefahr mit sich bringen kann, den Nutzer, der eine elementare Auskunft erwartete, mit etlichen speziellen Bedeutungen zu überschütten.

Ein Ausweg könnte in dem Verfahren liegen, wie es ABBY LINGVO gegenwärtig praktiziert, indem die allgemeinsprachliche Grundbedeutung eines Eintrags angegeben und gleichzeitig darauf verwiesen wird, dass etliche weitere fachsprachliche Sonderbedeutungen bereit stehen, die – nun wieder getrennt nach Fachgebieten – per Mausklick unmittelbar an Ort und Stelle verfügbar sind. Sicher nicht gangbar ist der Weg, den Mul-

tilex beschreitet, indem wiederum dem Nutzer überantwortet wird, das "richtige" Wörterbuch zu nutzen – völlig korrekt dabei wieder bei LINGVO, das sofort und selbstständig ins richtige Wörterbuch führt.

Doch wenn wir uns gegenwärtig angebotene elektronische Wörterbücher des Russischen (und nicht nur des Russischen) anschauen, verraten zudem viele von ihnen auch in der Bildschirmdarstellung häufig noch die Druckfassung als ihre ursprüngliche Quelle, indem sie den Nutzer mit kryptischen Abkürzungen, Querverweisen und "vergleiche-auch"-Ratschlägen nerven und auf ihrem eigenen Aufbauprinzip beharren, das man erst einmal erschließen muss. Sehr deutlich ist das beispielsweise in der elektronischen Fassung des von Frau Professor Belentschikow herausgegebenen, fachlich-sprachlich ausgezeichneten "Russisch-Deutschen Wörterbuchs" der Akademie der Wissenschaften und der Literatur Mainz, das auf Schritt und Tritt mit Abkürzungen um sich wirft, die drucktechnisch vollkommen berechtigt sind, in einer Bildschirmdarstellung aber befremden müssen – man vergleiche beispielsweise den wahllos herausgegriffenen Eintrag für "Maße, Abmessungen" – russisch "габари́т <m, GSg -a, gew. Pl>" – muss denn das so dargestellt werden, dass man sich hieraus "m = Maskulinum", "GSg -a = Genitiv Singular auf -a", "gew. Pl = tritt vorzugsweise im Plural auf" selbstständig erschließen muss? Doch auch Nachschlagewerke, die in gewisser Weise den Standard mit bestimmen, wie etwa die aus der PC-Bibliothek-Reihe, die auch die einschlägigen Ausgaben des Duden beinhalten, sind nicht abkürzungsfrei und können durch Weiterverweise schon zu Verunsicherungen führen, wenn beispielsweise beim Stichwort "Haare" (schön, dass die Pluralform überhaupt erfasst ist) auf "Haar" verwiesen wird, sich dieser Weiterverweis seinerseits dann in zwei Begriffe aufspaltet und außer der phraseologischen Einbindung nicht so ganz klar wird, was denn die beiden Einträge nun voneinander unterscheidet.

Zu diesen beiden Komplexen, die aktuelle elektronische Wörterbücher sicher nicht ganz gerechtfertigt von ihren in Buchform erschienenen Eltern übernommen haben, gesellen sich aber gewiss noch etliche weitere Momente, die von einer Printfassung schlechterdings nicht erwartet werden können, die sich aber bei einer multimedialen Umgebung, wie wir sie beim Computereinsatz schon fast selbstverständlich erwarten, von vornherein aufdrängen. In allererster Linie denkt man sicher an die akustische Veranschaulichung einer lexikalischen Einheit, eine Forderung, der heute bereits zahlreiche elektronische Übersetzungshelfer nachkommen (auch das bereits mehrfach als positives Beispiel benannte LINGVO führt zumindest 5000 lexikalische Einheiten in korrektem Englisch an, auf der Deutsch-Strecke und ebenso beim Französischen, Italienischen und sogar beim Russischen muss das benannte Wörterbuch leider passen, was umso bedauerlicher ist, als dass es ohnehin keine Angaben zur Betonung, zum Wortakzent enthält und damit auf diesem für das Russische relativ komplizierten Gebiet sogar noch einen Schritt hinter den typischen Printfassungen zurückbleibt). Ebenso sollte man erwarten können, dass Besonderheiten einer jeweiligen lexikalischen Einheit erläutert werden – sei es auf morphologischem paradigmatischem Gebiet, sei es hinsichtlich bestimmter Bedeutungsnuancen oder Gebrauchsbedingungen, sei es bei Aussprachebesonderheiten, sei es bei assoziativen Elementen, die sich mit der jeweiligen lexikalischen Einheit im Bewusstsein der Sprecher verknüpfen, oder einfach hinsichtlich der Erläuterung eines Sachhintergrunds, der mehr oder weniger enzyklopädischen Charakter trägt – der Autor des Wörterbuchs weiß ja nicht, warum der Nutzer gerade ein bestimmtes Wort nachschlägt, und sollte für alle diese Individualitäten und noch einige weitere gerüstet sein.

Kann das unbescheiden "Universalwörterbuch" genannte Hilfsmittel für das Sprachenpaar Russisch-Deutsch einen Schritt voran bedeuten und einige der monierten Schwächen abstellen? Schlüssig können die Autoren diese Frage sicher selbst nicht beantworten, zum einen, weil sie voreingenommen sind, zum anderen, weil dieses Wörterbuch zwar in einer Grundversion vorliegt, jedoch etliche geplante Funktionalität noch ihrer Umsetzung harrt.

2 Zielstellungen des RUW

Das **R**ussische **U**niversal**w**örterbuch – **RUW** – ist, sicher im Gegensatz zu manch anderen Wörterbüchern, von vornherein als elektronisches

Medium konzipiert worden. Es soll mit seiner Fertigstellung (gegenwärtig sind rund 130 000 Einträge grob erfasst, also darstellbar, aber noch nicht in allen Richtungen aufbereitet) den Wortschatz des aktuellen Russischen in drei Richtungen beschreiben:

a) soll es in seiner zuerst in Angriff genommenen und daher auch am weitesten gediehenen **didaktischen** Ausrichtung dem Lernenden Auskunft insbesondere zur morphologischen Beschaffenheit jeweiliger Wörterbucheinträge geben. Hierunter ist in erster Linie das paradigmatische Verhalten des Lexikon-Eintrages zu benennen, das bei *allen* flektierbaren Wörtern *alle* vorhandenen Wortformen vollständig ausgeschrieben anführt und etwaige Besonderheiten kommentiert. Für den Lerner wichtige Auskünfte etwa zu Frequenz und Distribution, zu Valenz bzw. Kombinatorik, darunter auch zu phraseologischen Bindungen, zu synonymischen, paronymischen und ggf. antonymischen Beziehungen sowie zu Wortverwandtschaften bzw. zur Etymologie stehen auf Abruf bereit. Mit genannten Stichwörtern ist im Prinzip die Integration von Einzelwörterbüchern in ein Gesamtwörterbuch gemeint, die isoliert schon existieren, aber unter didaktischem Aspekt noch nicht in derart komplexer Form zusammengeführt wurden, aber gerade bei der Wortschatzaneignung ist das Abstützen einer neu zu erlernenden lexikalischen Einheit durch bereits bekannte sprachlich verwandte, der Verweis auf und die Abgrenzung von Synonymen oder Antonymen oder die Warnung vor denkbaren Verwechslungsgefahren von immenser Bedeutung. Der Nutzer kann auf die Lemmata per Stichworteingabe zugreifen, aber auch Hyperlinks werden ausgiebig genutzt – praktiziert wird in breitem Maße beispielsweise das Einblenden der deutschen Bedeutung einer lexikalischen Einheit in entsprechende Materialien des Gesamtlehrwerks "Russisch aktuell" mit seinen stärker linguistischen Teilen ("Leitfaden" als theoretische Grammatik, "Phonetik" als Beschreibung der phonetisch-artikulatorischen und intonatorischen Gegebenheiten des Russischen) und den stärker auf die sprachpraktische Aneignung ausgerichteten Teilen ("Sprachkurs", "Sprechkurs", "Sprechendes Wörterbuch"). Innerhalb des computergestützten Fremdsprachenlernens bietet damit das RUW die lexikologisch-lexikografische Bezugsbasis für alle weiteren Teilkomponenten des Gesamtwerks, und in bestimmtem Maße ist es selbst direkt für didaktische Zielstellungen nutzbar (Einheiten aus dem RUW können beispielsweise zur Pflichtlexik erklärt und in russisch-deutschen wie deutsch-russischen "Vokabellisten" trainiert werden, bei denen Sprachrichtung, Ordnungsprinzip der Darstellung und Soundbegleitung über Voreinstellungen beliebig wählbar sind). Der oben beschriebene Suchmisserfolg soll insbesondere dem noch wenig erfahrenen Nutzer möglichst erspart bleiben, weswegen die Möglichkeit einer fehlertoleranten Eingabe vorgesehen wurde (bestimmte Orthografiefehler, die meist auf Interferenzen aus dem Deutschen zurückgehen, lassen sich mit großer Wahrscheinlichkeit prognostizieren, sodass entsprechende Umleitungslemmata den Lernenden behutsam in die richtige Richtung stupsen. Auch Eigennamen gehören zum Lexembestand des RUW – ihre Umlautung ist oft mit nicht geringen Verstehensproblemen verbunden, ihre morphologische Darstellung nicht unkompliziert. Die didaktische Ausrichtung impliziert auch, dass zumindest für den Grundwortschatz (bis zur Häufigkeit 10 000) ein Hörmuster bereitgestellt wird, und legt ebenso die Wiedergabe der lexikalischen Einheiten in phonetischer Umschrift nahe. Wenn dieses Kriterium der Häufigkeit angelegt wird, muss das RUW selbst mit Häufigkeitsangaben hantieren können. Schließlich sollte das RUW gerade unter seiner didaktischen Zielstellung als Formenwörterbuch auch Aussagen zur Formenbildung lexikalischer Einheiten treffen können, die nicht selbst im Wörterbuch erfasst sind, deren paradigmatisches Verhalten aber dank der Modellierung der morphologischen Varianten des Russischen mit genügender Sicherheit algorithmisch vorhergesagt werden kann. Durch dieses Prinzip der vollständigen paradigmatischen Auflistung ist das RUW mitunter zu Aussagen gezwungen, um das sich manch anderes Wörterbuch herummogelt – hat nun die russische Entsprechung

für den deutschen "Zucker" Pluralformen oder nicht? Eben diese Konsequenz trifft auch auf die Modellierung der Aussprache-regularitäten zu – Umschriftmuster können ebenfalls weitgehend algorithmisch gebildet werden und umfassen stets das gesamte Lemma, nicht nur die vermeintlich problematische Wortposition, wie es beispielsweise das Akademiewörterbuch handhabt.

b) soll das RUW in seiner **translationsbezogenen** Komponente dem Übersetzer helfen, in möglichst kurzer Zeit den jeweils treffendsten zielsprachlichen Ausdruck bei der Wiedergabe russischer Textausschnitte zu finden. Für den Einsatz in der übersetzerischen Praxis werden sich sicherlich einige der oben bereits benannten Informationen wiederfinden (auch der Übersetzer ist froh, wenn er vor "falschen Freunden" gewarnt wird oder unmittelbaren Zugriff auf eine phraseologische Einbindung hat). Neben den für das Übersetzen dominanten Vorschlägen für deutsche Entsprechungen, die sicher auch durch russische Minikontexte zu illustrieren sein werden, wird es vor allen Dingen um eine sehr exakte Beschreibung der stilistischen Besonderheiten und Restriktionen gehen. Für die Rationalität der übersetzerischen Arbeit wäre jegliche Limitierung des erfassten Wortschatzes von Nachteil; das Anführen von Neologismen ist zwingend. Für die übersetzerische Tätigkeit wäre das Nachschlagen auch flektierter Formen aus einer Textverarbeitung heraus günstig; die Möglichkeit der Wortformeneingabe wurde auch anfangs bereits als Desideratum benannt, wir kehren unten darauf noch einmal zurück. Das Ergänzen des Wörterbuchoriginals durch benutzereigene Einträge ist sicher wünschenswert.

c) können die vorhandenen vielfältigen Informationen, die das RUW zum Russischen enthält, auf einer **linguistischen** Ebene nutzbar gemacht werden, indem für linguistische Untersuchungen über bestimmte Abfragemechanismen alle lexikalischen Einheiten mit gerade gesuchten morphologischen, stilistischen oder anderen Besonderheiten herausgefiltert werden. Solche Abfragemechanismen sollten sowohl die Lexikoneinträge und ihre Spezifika selbst wie auch die gebildeten Wortformen betreffen; ihre Realisierung ist bei der Konzipierung des Wörterbuchs berücksichtigt, indem ein und dieselbe Besonderheit stets auf programmtechnisch wieder auffindbare Art und Weise kommentiert wurde. Die Auflistung der Lemmata in rückläufiger Sortierung ist möglich und hat bereits bei der Erarbeitung wertvolle Dienste geleistet. In vollem Umfang wird sich diese sprachwissenschaftliche Recherchemöglichkeit erst ergeben, wenn alle geplanten Einträge aufgenommen und sämtliche Wortformen zu sämtlichen Einträgen generiert werden können (diese Generierungsmöglichkeit funktioniert jetzt schon, ist aber noch nicht sinnvoll einsetzbar, da die Menge der Grundeinträge noch zu niedrig liegt). Die komplett generierten Wortformen sollen später einmal Eingang in eine Vollwortformen-Datenbank finden, die auch für die Stichworterkennung eingesetzt werden kann.

Genannte Funktionalität, die natürlich nur in einer computergestützten Fassung erreichbar ist, liegt weit über der vergleichbarer anderer Wörterbücher; dabei ist es durchaus nicht ketzerisch, die Frage zu stellen, ob eine solche Breite der Anlage einem tatsächlichen Praxisbedürfnis entspricht. Wir bejahen diese Frage in dem Sinne, dass eine Beschreibung der lexikalischen Einheiten einer Sprache so umfassend vorgenommen werden sollte, wie es vorn dargelegt wurde, dass es aber gleichzeitig möglich sein muss, bei der Präsentation der Informationen zu dieser lexikalischen Einheit in unterschiedliche Tiefen der Beschreibung vorzustoßen und auch unterschiedliche Oberflächendarstellungen zu ermöglichen, denn es ist in der Tat so, dass der gestandene Übersetzer, der das RUW vorrangig als russisch-deutsches Nachschlagewerk nutzen will, an einer Paradigmendarstellung nur sehr bedingt interessiert ist, ebenso wie der Russischanfänger, der sich das Verb "fahren" erschließt, dessen phraseologische Einbindung oder umgangssprachliche Nebenbedeutungen vorerst herzlich kalt lassen. Ebenso ließe sich vermuten, dass nur der Lernanfänger an einer textuell eingebundenen Kommentierung der morphologischen Besonderheiten einer lexikalischen Einheit interessiert ist, während der erfahrenere Nutzer mit einer sehr viel

stärker komprimierten, auf wenige Stichworte reduzierten Darstellung einen ebensolchen Informationsgewinn erzielen kann. Ähnlich wie bei der theoretischen Grammatik, die Bestandteil des Gesamtwerks ist, soll also auch das Russische Universalwörterbuch auf spezifische Nutzerinteressen und –bedürfnisse eingehen und eine immer größere Spezifizierung der zu gewinnenden Erkenntnisse ermöglichen können, seine Daten also auf ganz unterschiedliche Weise präsentieren können – eine Aufgabe, die erst in Ansätzen gelöst ist, denn schon so stellt sich die Erarbeitung eines derart breit angelegten Nachschlagewerks als nicht gerade einfach dar.

3 computerlinguistische Umsetzung

Bei einem Universalwörterbuch haben wir es mit einer Datenmenge erheblicher Größenordnung zu tun. Eine solche auf zeitgemäßen PCs zu verwalten, ist zwar heute in Bezug auf Speichergröße und Zugriffszeit eine beherrschbare Aufgabe, diese Datenmenge jedoch zu erstellen, zu pflegen, zu erweitern und ihre Richtigkeit abzusichern, ist mit außerordentlich hohem personellen Aufwand verbunden, woraus sich als Grundforderung für uns ableitete, manuelle Tätigkeiten dieser Art auf das erreichbare Minimum zu reduzieren und möglichst weitgehend rechentechnisch zu unterstützen.

3.1 Datenbasis

Unser Grundprinzip bei den Wörterbuchdaten ist es, in der Datenbasis des RUW nur solche Informationen explizit zu führen (also manuell erfassen zu müssen), die nicht algorithmisch aus dem Zeichenkörper des Lemmas selbst oder aus anderen explizit bereits vorhandenen Informationen abgeleitet werden können. Entsprechend hoher Aufwand bei der programmtechnischen Realisierung ist die logische Folge. So werden beispielsweise alle Wortformen eines flektierbaren Lemmas durch den integrierten algorithmischen Formengenerator bereitgestellt (singuläre Ausnahmen müssen natürlich explizit erfasst sein). Die morphologische Analyse des Zeichenkörpers eines Lemmas erlaubt meist, Annahmen über die Wortart zu treffen, das anzuwendende Paradigma vorherzusagen und auch bestimmte Besonderheiten im Flexionsverhalten zu berücksichtigen, die sich teils ausschließlich aus dem Zeichenkörper ableiten lassen, teils auf explizite Zusatzinformationen stützen müssen. Die vom morphologischen Analysator gewonnenen Informationen werden einerseits für die Synthese der Flexionsformen verwendet, sie können andererseits auch einzeln in Hypertexten abgefragt werden, um entsprechende sprachliche Erscheinungen für den Lerner bzw. Wörterbuchbenutzer ausführlich zu kommentieren.

Da recht viele verschiedene Informationen in einem Wörterbucheintrag vorkommen können (ca. 250 verschiedene werden derzeit unterstützt), die Anzahl der bei einem konkreten Wörterbucheintrag vorkommenden expliziten Informationen aber – auch dank der algorithmischen Bereitstellung – eher gering ist, bietet sich das Schlüsselwortformat für die Speicherung an, wobei eine weitgehende Formalisierung der Einträge angestrebt wurde, um die Daten effektiv auffinden, prüfen und später auch recherchieren zu können. Jedoch sind auch rein textuelle Anmerkungen möglich, um den Autoren der Wörterbucheinträge auch für seltenste sprachliche Absonderlichkeiten Kommentierungsmöglichkeiten in die Hand zu geben.

Der Umfang der Wörterbucheinträge variiert stark. Jedenfalls angegeben ist die Wörterbuchform selbst mit Angabe der Wortakzente (\acu bei Festbetonung bzw. \gra bei Wechselbetonung) und die deutsche Bedeutung:

[**авт**\acu **обусный**]
D=Autobus-, Bus-_ *deutsche Bedeutung*
AB1=3 *Ableitungsbasis автобус(3 Zeichen kürzer)*

Bei Eigennamen kann jedoch ggf. automatisch die Umschrift als „deutsche Bedeutung“ generiert werden:

[\acu **Африка**]
D=^ *automatische Umschrift*
FW=E *Elementarwortschatz*
W=SUO *Substantiv, Eigenname/Ortsname*
AB=!GGS *Hinweis: Großschreibung bei Eigennamen*

Diese direkte Übernahme bietet sich insbesondere bei Ortsnamen, Vor- und Familiennamen oder Benennungen mit Eigennamencharakter an. Bei Verben kann die deutsche Bedeutung ganz oder teilweise vom Aspektpartner bezogen werden:

[**пис**\gra **ать**]
KL=11 *Verb-Klasse*
PP=напис\gra ать *perfektiver Partner*
D=schreiben; malen; Musik (im Tonstudio) aufnehmen
FW=E *Elementarwortschatz*
AG=!NOG *Hinweistext zu gemiedenen Formen*
FQ=199;578 *Frequenz: Rang, Zählwert*

[**напис**\gra **ать**]
KL=11
IP=пис\gra ать *imperfektiver Partner*
FW=E
FQ=419;303

Man sieht bei den gezeigten Einträgen, dass nur ein Minimum an Informationen erfasst wird. Die knappe Informationsmenge gestattet es jedoch, sowohl alle Flexionsformen zu generieren als auch Aussagen über eine Reihe von Besonderheiten zu machen.

Der Zugriff auf die Daten des Wörterbuchs basiert auf einem Hashcode, der aus dem Lemmanamen bestimmt wird. Durch diese Zugriffsmethode ist ein außerordentlich schneller Zugriff auf die Lemmata möglich, ohne dass das gesamte Wörterbuch in den Speicher geladen werden muss. Ein Cashing für bereits in den Hauptspeicher übernommene Lemmata sorgt zusätzlich für einen Performance-Gewinn, der noch dadurch erhöht wird, dass einige häufig benutzte Informationen, die extern ggf. nur implizit vorliegen, beim Laden in den Cache als explizite Daten im Hauptspeicher bereitgestellt werden.

3.2 Formengenerierung

Für die einigermaßen regulären Formen sind keinerlei Angaben in der Datenbasis erforderlich. Mit einer Reihe von Autodetect-Funktionen z. B. für die Wortart, den Betonungstyp, die Subklassifizierung bei Adjektiven und Substantiven, den Deklinations- bzw. Konjugationstyp u. a. werden diese Eigenschaftswerte so ermittelt, dass nur in den Ausnahmefällen, die natürlich immer noch zahlreich genug sind, explizite Informationen

Paradigma für писàть **(1.1)**

Das transitive Verb **писàть** [schreiben; malen; Musik (im Tonstudio) aufnehmen] ist imperfektiv, perfektiver Partner ist написàть.

Wie das Musterverb вязàть ist писàть (§ 90.3) ein unproduktives Verb auf -ать, I. Konjugation (= e-Konjugation) 1. unproduktive Verbalgruppe. Bei diesen Verben ist (durchgängiger) Konsonantenwechsel zwingend:

Beim Stammauslaut von писàть (-с-) tritt **Konsonantenwechsel zu -ш-** auf. Ein Betonungswechsel ist nach AT 7 sehr wahrscheinlich; писàть ist daher **wechselbetont**.

Präsensformen: NOG		Präteritalformen:	
Stamm:	пиш-	Stamm:	писа-
я	пишỳ	он	писàл
ты	пѝшешь	она	писàла
он, она	пѝшет	оно	писàло
мы	пѝшем		
вы	пѝшете	они	писàли
они	пѝшут		
Imp. Sg.:	пишѝ!	P. Passiv:	напѝсанный
Imp. Pl.:	пишѝте!	KF m.:	напѝсан
		KF f.:	напѝсана
P. Passiv:	-	KF n.:	напѝсано
		KF pl.:	напѝсаны
P. Aktiv:	пѝшущий	P. Aktiv:	писàвший
AdvP. ipf.:	-	AdvP. pf.:	написàв

benötigt werden.

Notwendige Anpassungen der Orthografie, Einfluss von Palatalisierung/Jotierung, und zahlreiche andere Anpassungen wie Konsonantenwechsel, e- oder o-Ausfall bzw. -Einschub oder Betonungswechsel werden bei der Analyse des Zeichenkörpers erkannt oder zumindest mit einiger Wahrscheinlichkeit vermutet und bei der algorithmischen Formengenerierung entsprechend berücksichtigt. Diese Grundvermutungen werden auch bei der Behandlung von Wörtern herangezogen, die im Lexikon nicht vorhanden sind, mit den entsprechenden einschränkenden Hinweisen aber dennoch in ihrer Formenbildung dargestellt werden können.

Dennoch sind zahlreiche Ausnahmen zu berücksichtigen. Irregularitäten bei der Formenbildung gibt es in verschiedenem Grade:

a) Ein Wort verhält sich nicht so, wie man es anhand seines Zeichenkörpers (insbesondere am Wortende) vermuten würde, es ist jedoch lediglich einer anderen Flexionsklasse zuzuordnen. Die manuelle Angabe der korrekten Klasse löst dieses Problem.
b) Der von der Analyse des Infinitivs ermittelte Wortstamm ist nicht geeignet, alle Formen daraus abzuleiten; einigen oder allen Formen liegt ein irregulär modifizierter Stamm (wie bei де́рево im Plural дерѐвья) oder ein gänzlich anderer Stamm (wie bei человек und люди) zugrunde. In diesen Fällen genügt es meist, den jeweiligen produktiven Stamm explizit anzugeben. Bei Verben können bis zu 5 modifizierte Stämme eine Rolle spielen.
c) Bestimmte Gegebenheiten lassen mit einiger Wahrscheinlichkeit auf die Anwendbarkeit bzw. Nichtanwendbarkeit von Sonderbehandlungen schließen, das konkrete Lemma verhält sich aber gerade anders als die Mehrheit der diesbezüglich vergleichbaren Lemmata. In diesen Fällen müssen zusätzliche Eigenschaften wie "e- oder o-Einschub ist doch nicht anwendbar" oder "Ein anderer als der vermutete Betonungstyp ist anzuwenden" u. a. entsprechend formalisiert angegeben werden.
d) Einzelne Formen entziehen sich völlig der algorithmischen Behandlung und müssen deshalb explizit angegeben werden. Dazu gibt es für jede denkbare grammatische Form ein entsprechendes Schlüsselwort.

Eine Kombination der Fälle c) und d) liegt beispielsweise bei стул vor:

[**стул**]
PST=0:2ьj *Pluralstamm*
P1=ст\acu ул2ь3я *1. Person Plural*
P2=2\acu ул2ь3ев *2. Person Plural*
...

Im Plural lassen sich Dativ, Instrumental und Präpositiv von стул anhand des mit PST angegebenen Pluralstamms algorithmisch bilden, Nominativ und Genitiv jedoch nicht. In den angegebenen Formen sind Ziffern enthalten, die zwei verschiedenen Zwecken dienen: Eine am Anfang angegebene Zahl ist eine sogenannte Rückschnittzahl. Sie kann benutzt werden, um eine Referenzzeichenfolge (hier der Lemmaname стул) zu übernehmen, von der rechts entsprechend viele Zeichen zu tilgen sind. Bei PST steht die 0 also für стул. Der Doppelpunkt ist ein Trennzeichen, das verwendet werden muss, wenn der Rückschnittzahl Ziffern folgen, womit wir bei der Bedeutung der übrigen Ziffern sind: Sie stehen als Tags in den Formen und werden verwendet, um Endungen, Suffixe bzw. modifizierte Zeichen des Stamms in der Anzeige farblich hervorgehoben darstellen zu können. Auch die algorithmisch generierten Formen enthaltendiese Tags. Bei P2 ist die erste "2" wiederum eine Rückschnittzahl, deren Speicher und Tipparbeit sparende Wirkung bei so kurzen Wörtern wie стул in unserem Beispiel nicht ersichtlich ist (2\acu ул2ь3ев ist äquivalent zu ст\acu ул2ь3ев).

Gelegentlich existieren konkurrierende Formen, sodass neben einer (meist regulären, algorithmisch erzeugbaren) Form ein oder selten mehrere weitere Formen denkbar sind. Solche alternativen Formen sind in die Datenbasis einzupflegen:

[**драг**\acu **ун**]

P2=+драг\acu ун
...

(im Genitiv Plural ist sowohl das generierte драгунов als auch das zusätzliche драгун möglich). Derartige Alternativangaben können in verschiedener Form auftreten:

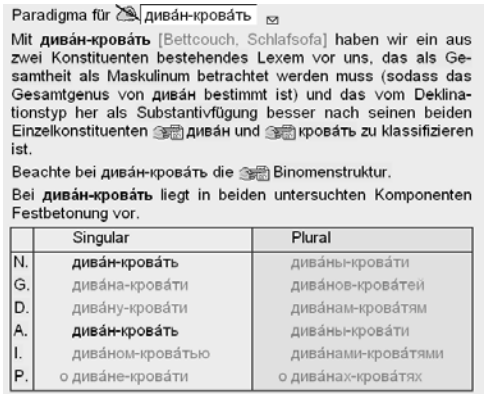
+<form> *zusätzlich, nach regulärer Form*
<form>+ *vor regulärer Form anordnen*
<form>**(**+ *reguläre Form in Klammern anschließen*
<form>**(@**+ *(alt: <reguläre Form>) dahinter*

Ein weiteres Problem stellen Formen dar, die von bestimmten Lemmata nicht bildbar sind oder gemieden werden. Bei diesbezüglich regulären Lemmata kann in Abhängigkeit von bestimmten anderen Eigenschaften die Abdeckung der wortartspezifischen Formenmenge vorhergesagt werden: Bei Verben hat beispielsweise der Aspekt Einfluss auf die Formenmenge, bei Adjektiven hilft die Subklassifizierung, um feststellen zu können, ob Komparativ, Superlativ bzw. Kurzformen existent sind, bei Substantiven deuten verschiedene Suffixe auf Numerusdefektivität hin (ähnlich wie bei deutsch -heit oder -keit sind russische Wörter auf -ость bzw. -ние oft, aber eben nicht immer Singulariatantum). In anderen Fällen gibt es semantisch motivierte oder andere Gründe, dass bestimmte Gruppen von Formen nicht gebildet werden. Die Generierung dieser Formen kann durch
NOF=<formengruppenname>
unterbunden werden:

[**ввод**\gra **иться**]
D=in Gebrauch kommen, sich einbürgern, üblich werden
NOF=12 *Die Formen der 1. und 2. Personen fehlen*
...

In zahlreichen Fällen werden dabei automatisch entsprechende textuelle Hinweise zu den ausfallenden Formen bereitgestellt. Falls zum konkreten Lemma spezielle Hinweise zum Formendefizit angeraten erscheinen, können diese – wie viele andere Hinweise auch – als Sondertexte oder unter Verwendung vorgefertigter Textblöcke vom Bearbeiter des Lemmas in der Datenbasis bereitgestellt werden.

Um die Menge der notwendigen Angaben weiter zu reduzieren, wurde die Möglichkeit geschaffen, Angaben, die bei einem anderen


Paradigma für диван-кровáть

Mit **диван-кровáть** [Bettcouch, Schlafsofa] haben wir ein aus zwei Konstituenten bestehendes Lexem vor uns, das als Gesamtheit als Maskulinum betrachtet werden muss (sodass das Gesamtgenus von **диван** bestimmt ist) und das vom Deklinationstyp her als Substantivfügung besser nach seinen beiden Einzelkonstituenten **диван** und **кровáть** zu klassifizieren ist.

Beachte bei **диван-кровáть** die Binomenstruktur.

Bei **диван-кровáть** liegt in beiden untersuchten Komponenten Festbetonung vor.

	Singular	Plural
N.	**диван-кровáть**	диваны-кровáти
G.	дивана-кровáти	диванов-кровáтей
D.	дивану-кровáти	диванам-кровáтям
A.	**диван-кровáть**	диваны-кровáти
I.	диваном-кровáтью	диванами-кровáтями
P.	о диване-кровáти	о диванах-кровáтях

ähnlichen Lemma bereits vorhanden sind, mit nutzen zu können. Mit der Angabe R=<lemmaname> müssen nur noch alle diejenigen Angaben aufgeführt werden, die gegenüber dem R-Lemma unterschieden sind.

Adverbien müssen nur dann erfasst werden, wenn sie nicht streng mit einem Adjektiv assoziiert sind oder sich durch Besonderheiten von diesem abheben; ansonsten werden alle benötigten Eigenschaftswerte automatisch aus den Einträgen der in Beziehung stehenden Adjektive entnommen.

Wenn auch sehr viele Eigentümlichkeiten der Wortformenbildung mit hoher Zuverlässigkeit vorhergesagt und damit auch modelliert werden können, liegt der Anteil archaischer und damit irregulärer Formen nicht niedrig; jeder, der sich einmal mit dem Erlernen von Fremdsprachen befasst hat, wird beispielsweise die Crux der irregulären Verben aus eigener Erfahrung kennen. Eine Effektivierung der Wörterbucheinträge wird bei diesen Verben dadurch erreicht, dass für sie in ihren unterschiedlichen Präfigierungen die morphologischen und übergreifenden Eigenschaften nur einmalig bei einem jeweiligen Pseudolemma erfasst werden, auf das alle durch Präfigierung entstandenen Abkömmlingen zurückgreifen. Dieser Mechanismus ist deshalb von besonderem Wert, weil gerade die hochfrequenten irregulären Verben (die zahlreiche Ausnahmekennzeichnungen und damit umfangreiche Einträge haben) mit einer ausufernden Anzahl von Präfixen Verbindungen eingehen; die Zahl der für das Russische aufgestellten unterschiedlichen Subklassen hat die Hundert überschritten, die Zahl der Ableitun-

gen wird sich in der Größenordnung zwischen 2 000 und 3 000 (bei ca. 18 000 zur Zeit erfassten Verben) bewegen.

Mehrworttermini

Mehrworttermini, Begriffe, die aus zwei oder mehr Konstituenten bestehen, sind im Russischen vergleichsweise häufiger als im Deutschen, da eine Kompositabildung fehlt – so ist eine Schlafcouch ein диван-кровать und eine Bushaltestelle eine автобусная остановка.
Solche Fügungen sind zumindest hinsichtlich ihrer deutschen Bedeutung zu behandeln, aber auch Hinweise zu ihrer Strukturierung sind häufig angebracht. Deshalb führt das RUW Mehrworttermini in großer Zahl an. Bezüglich der grammatischen und morphologischen Eigenschaften kann dabei die überwiegende Mehrzahl der Eigenschaften von den einzelnen Konstituenten übernommen werden. Sowohl bei der Kommentierung als auch bei der algorithmischen Komposition der Flexionsformen von Mehrwortterminі muss allerdings die Struktur des Lemmas berücksichtigt werden. Dazu werden im RUW derzeit 15 strukturelle Grundtypen unterschieden. Der Grundtyp wird explizit geführt; er impliziert Annahmen für die Position des Basisworts sowie die Anordnung und Valenz der übrigen Konstituenten. Allerdings kann das Strukturmodell eines Mehrwortterminus‘ auch explizit formalisiert angegeben werden, um auch seltenere Strukturen von Mehrworttermini behandeln zu können. Da die Konstituenten einerseits als Flexionsformen im Mehrwortterminus vorkommen können und andererseits ggf. eine Disambiguierung der Referenzen auf die Konstituentenlemmata möglich sein muss, können die Lemmanamen der beteiligten Konstituenten explizit angegeben werden, falls ihre Schreibung innerhalb des Mehrwortterminus‘ von Ihrem Lemmanamen abweicht:

[**пр**\gra **аво г**\gra **олоса**]
D=Stimmrecht, Wahlrecht; Redefreiheit
V=g
K1=~ (на чт\gra о-н.) *~ für пр\gra аво*
K2=1 *Rückschnitt: г\gra олоса ohne das а*
Q=s *Singularetantum*

Homonymie und Polysemie

Zahlreiche Wörter des Russischen (wie wohl jeder anderen Sprache auch) unterscheiden sich im Zeichenkörper der Wörterbuchform nicht, obwohl sie mehr oder minder verschiedene Bedeutungen haben. Ist der Bedeutungsunterschied hinreichend groß oder gibt es gar grammatische oder morphologische Unterschiede, sind sinnvollerweise getrennte Lemmata im Wörterbuch zu führen. Beispiel пропасть:

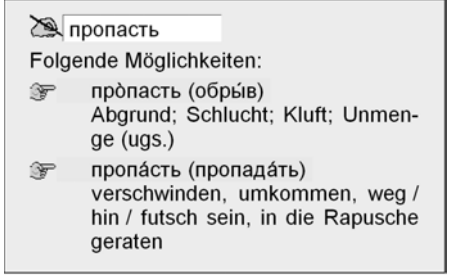
[**пр**\gra **опасть** (обр\acu ыв)]
D=Abgrund; Schlucht; Kluft; Unmenge (ugs.)
AQ=пропаст\acu и (в\acu ыпасти)
...

[**про|п**\acu **асть** (пропад\acu ать)]
D=verschwinden, umkommen, weg\~/ hin\~/
futsch sein, in die Rapusche geraten
IP=пропад\acu ать (проп\acu асть)
AQ=пропаст\acu и (в\acu ыпасти)
W=V
...

Beide Lemmata werden unabhängig voneinander beschrieben. Die Lemmanamen werden durch Hinzufügen von Klammerbemerkungen eindeutig gemacht. Gibt der Wörterbuchbenutzer nun пропасть ein, ist zunächst unklar, welches der beiden Lemmata anzuzeigen ist. Das RUW präsentiert deshalb eine Auswahlliste aller vorliegenden Homonyme und Polyseme des eingegebenen Wortes. Bei der Anzeige des ausgewählten Lemmas wird automatisch ein Hinweis bereitgestellt, dass es mehrdeutige Einträge gibt, die dann über ein Popup angesteuert werden können.

3.3 Präsentation der Lemmata

Die zu einem Lemma darzustellenden Informationen sollen den eingangs dargestellten Forderungen entsprechend in übersichtlicher und gut lesbarer Form dargestellt werden. Dazu wird der am URZ der Uni Leipzig entwickelte Hypertext-Viewer **HyView** verwendet. Etwa 20 sehr variable Hypertextseiten, die die Schnittstelle zum RUW nutzen und die durch das anzuzeigende Lemma parametrisiert werden, dienen der Präsentation der Lemmata, weitere ca. 900 ebenfalls umfangreich parametrisierte Seiten mit Zusatzinformationen sind ausgehend von der Lemma-Präsentation erreichbar. Zahlreiche Links gestatten bei entsprechendem Erklärungsbedarf den



sofortigen Wechsel zu systematischen Darstellungen sprachlicher Phänomene in den Leitfaden des Russischen oder auch zu Erklärungen, die im Sprachkurs gegeben werden, der zudem außerordentlich umfangreiches Übungsmaterial bereithält; umgekehrt ist aus diesen Materialien heraus der uneingeschränkte Zugriff auf paradigmatische Darstellungen und Erläuterungen des RUW möglich. Für die automatische Lösungskontrolle wird ebenso in starkem Maße die Funktionalität des RUW herangezogen, mit dessen Hilfe es auch möglich ist, dem Übenden hilfreiche Kommentierungen der von ihm gemachten Fehler sowie effiziente Nachschlagemöglichkeiten anzubieten.

Für die Darstellung der Lemmata werden eine Reihe von Features genutzt, deren detaillierte Beschreibung den Rahmen dieses Artikels sprengen würde, deshalb seien einige wichtige hier nur aufgezählt:

- Sound-Einbindung für den gesamten Grundwortschatz des Russischen
- automatische Transliteration/Transkription vom Kyrillischen ins Lateinische unter Beachtung der verschiedenen Transkriptionssysteme für französisch, deutsch, englisch, bilbliothekarisch, wissenschaftlich, GOST, Duden usw.)
- automatische phonetische Transkription (IPA sowie Avanesov)
- automatische Silbentrennung für Deutsch und Russisch sowie die zugehörigen Tools zur Absicherung der durchgehenden Korrektheit der potenziellen Trennstellen.
- tief geschachteltes System von Textblöcken mit Substitutionen
- Makros und Funktionen, die in die Hypertext-Sprache RRT eingebettet sind.

3.4 Einträge editieren und prüfen

Die Erfassung, Korrektur und Erweiterung der Lemmata geschieht in unserem Fall durch mehrere Bearbeiter, die lokal voneinander getrennt arbeiten. Unterstützt wird diese Arbeitsweise durch die Möglichkeit, bearbeiterbezogene Änderungsdateien mit sofortiger Wirksamkeit der Änderungen in den vom RUW präsentierten Hypertexten zu führen, wobei die Änderungsdateien in gewissen Zeitabständen vom Hauptbearbeiter einer Endkontrolle unterzogen und relativ endgültig in den Datenbestand aufgenommen werden (spätere Korrekturen sind trotz allem noch möglich).

Dass sich beim Bearbeiten von Wörterbucheinträgen Fehler einschleichen, ist menschlich und wohl unvermeidbar; deshalb ist es von besonderer Wichtigkeit, alle Möglichkeiten der maschinellen Prüfung von Wörterbucheinträgen auszuschöpfen. Im Falle des RUW geschieht dies in vier Ebenen:

1. lemmabezogen: Einhaltung der Lemma-Namenskonventionen, Vollständigkeit der Minimalinformationen (z. B. deutsche Bedeutung, Wortart-Erkennbarkeit, ggf. erforderliche weitere Angaben, deren Notwendigkeit sich anhand des Zeichenkörpers des Lemmas oder aufgrund des Vorliegens anderer expliziter Informationen ergibt);
2. schlüsselwortbezogen: Plausibilitätskontrolle anhand des Wertevorrats bzw. der Syntax beim jeweiligen Schlüsselwort, Widerspruchsfreiheit mit anderen Schlüsselwortdaten desselben Lemmas,
3. referenzielle Integrität der Bezüge zu anderen Lemmata und der Links zu Erklärungen und Zusatztexten,
4. Hypertext-Kontrolle: Auffinden toter Links zu Lemmata des RUW in allen Teilen von "Russisch aktuell"

3.5 Kommentierende Angaben

Über die Beschreibung der deutschen Entsprechung und die Darstellung aller Flexionsformen sowie die Kommentierung ihrer Besonderheiten hinaus sind in der Anlage der Lemmadaten weiterführende Informationen vorbereitet, die einmal sichern sollen, dass das RUW seinem Anspruch, ein Universalwörterbuch zu sein, auch gerecht wird. Insbesondere meinen wir hier Angaben zur Paronymie (Verwechslungsgefahr von Wörtern tritt sowohl innerhalb des Russischen als auch durch Interferenz mit dem Deutschen auf), Synonymie (bei sorgfältiger stilistischer Differenzierung), Antonymie (die besonders unter didaktischem Aspekt im Sinne assoziativer Ketten von Bedeutung ist) und Etymologie (Aufzeigen von Wortverwandtschaftsbeziehungen unterstützt ebenfalls den Lerneffekt).

Informationen zur Phraseologie dürfen ebenfalls nicht ausgespart werden, unterscheidet sich doch oft die wörtliche Bedeutung eines Stichworts von seinem Auftreten bei der Einbindung in

feste oder sprichwörtliche Redewendungen. Teile dieser Darstellungen, die, wenn man sie für sich genommen betrachtet, Einzelwörterbücher im Wörterbuch darstellen, sind gegenwärtig bereits verfügbar, ihre endgültige Form wird sich jedoch erst bei der weiteren Arbeit am RUW herauskristallisieren.

3.6 Schnittstellen

Das RUW stellt seine Funktionalität an Schnittstellen bereit. Eine DLL für beliebige Programme, primär für Delphi oder C, gestattet den Zugriff auf alle expliziten sowie generierten Teilinformationen zu jedem Lemma. HyView, das Präsentationsprogramm für die Hypertexte der Lern- und Wörterbuchsoftware, bietet Funktionen und Makros, die auf der DLL aufsetzen und die sowohl einzelne als auch komplexe Informationen abrufen oder das Wörterbuch beim jeweils gewünschten Lemma aufschlagen können. Für das Recherchieren in den linguistischen Daten kann entweder mit den Funktionen der DLL oder mittels Delphi-Objekten zugegriffen werden, z. Z. wird diese Möglichkeit vorwiegend zur Qualitätssicherung der Wörterbucheinträge von den verschiedenen Prüffunktionen benutzt.

4 Zusammenfassung

Was ist das Besondere am RUW? Wohl die Breite der Gesamtanlage, der hohe Integrationsgrad von Wörterbuch, wissenschaftlicher Beschreibung des Russischen, Lehrmaterial und Übungen sowie der hohe Anteil computerlinguistischer Mechanismen von der morphologischen Analyse und Formensynthese über die automatische Silbentrennung bis hin zur Formengenerierung, von algorithmischer Transliteration und Transkription in andere Schriftsysteme bis hin zur maschinellen phonetischen Transkription.

Literatur

[1] *Abby Lingvo 8.0* Multilingual Edition, Abby Software House 2002

[2] *Multilex 4.0 German*, German-Russian Electronic Dictionary, MediaLingua JSC 2002

[3] Renate Belentschikow, *Russisch-Deutsches Wörterbuch*, Harrassowitz Verlag Wiesbaden 2003

[4] *Duden Band 8, Sinn- und sachverwandte Wörter*, Neudruck der 2. Auflage, Bibliographisches Institut & F.A.Brockhaus AG, Mannheim 1997

[5] Bernd Bendixen, Horst Rothe, Wolfgang Voigt, *Russisch aktuell – Der Leitfaden*, CD und Buch, Harrassowitz Verlag Wiesbaden 2003, ISBN 3-447-04816-6

[6] Bernd Bendixen, Galina Hesse, Horst Rothe, *Russisch aktuell – Der Sprachkurs*, CD und Buch, Harrassowitz Verlag Wiesbaden 2003, ISBN 3-447-04816-0