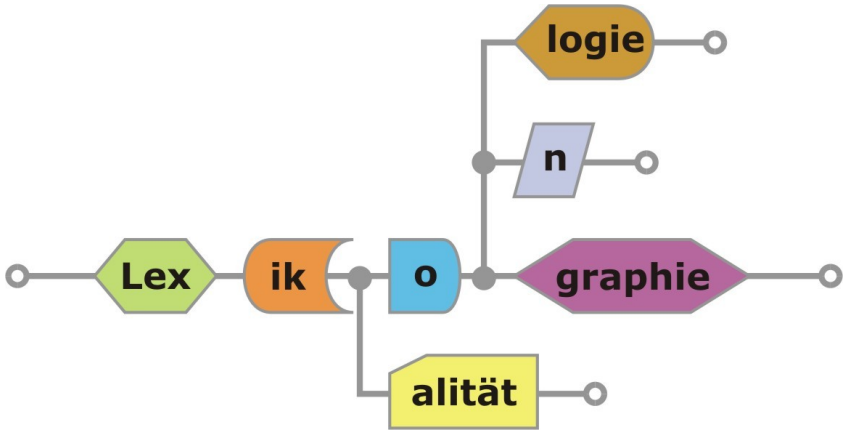
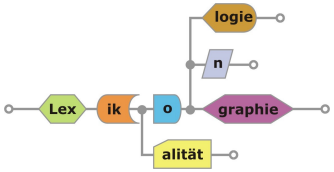
Bundesministerium
für Bildung
und Forschung

Wechselwirkungen zwischen linguistischen und bioinformatischen Verfahren, Methoden und Algorithmen:

Modellierung und Abbildung von Varianz in Sprache und Genomen

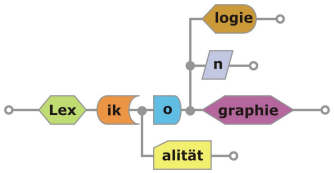


Dr. Andrea Rapp, Kompetenzzentrum Trier
www.kompetenzzentrum.uni-trier.de
andrea.rapp@uni-trier.de



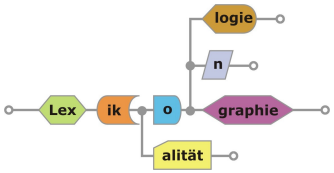
Inhalt

- Vorstellung des Förderprogramms & der Projektpartner
- Vorstellung der Projektsstruktur & der Zielsetzung
- linguistische Teilvorhaben: Metalemmaliste
- bio-informatisches Teilvorhaben: Erstellung von Ontologien, quantitative Analysen
- informatisches Teilvorhaben: integrativ
- Identifikation analoger Mechanismen
- Zielsetzung: Anwendungs- und Einsatzmöglichkeiten



BMBF-Förderprogramm "Wechselwirkungen zwischen Natur- und Geisteswissenschaften"

- In diesem Förderschwerpunkt arbeiten geistes- und naturwissenschaftliche Fächer in interdisziplinären Forschungsverbünden zusammen.
- In die interdisziplinäre Zusammenarbeit müssen sowohl geisteswissenschaftliche als auch naturwissenschaftlich-technische, mathematische oder informationstechnologische Kompetenz einfließen.
- Sowohl der Einsatz naturwissenschaftlicher Methoden in den Geisteswissenschaften als auch umgekehrt, der Einsatz geisteswissenschaftlicher Methoden in den Naturwissenschaften ist erwünscht, um nicht zuletzt neue Methoden zu entwickeln.



Vorstellung der Verbundpartner

Verbundkoordinator

Kompetenzzentrum
für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften

Prof. Dr. Claudine Moulin, Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften, Universität Trier

Verbundpartner

INSTITUT FÜR DEUTSCHE SPRACHE

Prof. Dr. Ludwig Eichinger, Institut für Deutsche Sprache, Mannheim

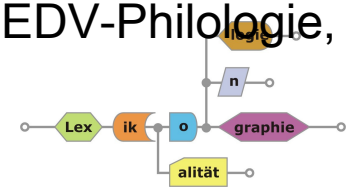
Prof. Dr. Dietmar Seipel, Institut für Informatik, Lehrstuhl für Informatik I, Informationsstrukturen und wissensbasierte Systeme, Universität Würzburg

BIOZENTRUM
UNIVERSITÄT
WÜRZBURG

Prof. Dr. Jörg Schultz, Dipl.-Biol., Theodor Boveri-Institut für Biowissenschaften/Bioinformatik, Universität Würzburg

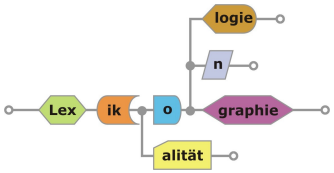
Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Prof. Dr. Werner Wegstein, Kompetenzzentrum EDV-Philologie, Universität Würzburg



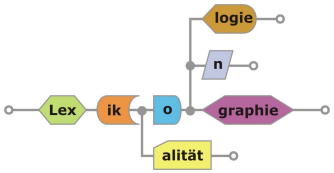
Fragestellung und Ziel

- prinzipielle strukturelle Gemeinsamkeiten
  - zwischen Genomcode und sprachlichem Code
  - biologischer Entwicklung und Sprachentwicklung
- zentrale gemeinsame Kennzeichen (u.a.)
  - Entwicklungsfähigkeit -> Historizität
  - Vielfalt bzw. Varianz der Erscheinungen
- differenzierteres Verständnis der Konzepte und Mechanismen von Entwicklung und Varianz ermöglicht
  - neue und präzisere Verfahren der Informationsgewinnung
  - der Speicherung
  - Bearbeitung und
  - Auswertung der dadurch gewonnenen Daten



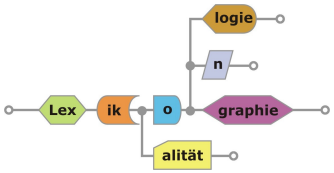
Projektstruktur: 4 Teilvorhaben

- TV1: Basis-Lemmaliste der neuhochdeutschen Standardsprache
- TV 2: Klassifizierte Varietäten-Lemmalisten, semasiologische Vernetzung, Erkundungsmodul onomasiologische Vernetzung
- TV3.1: Grammatik der Varianz in Genomen und Sprache, Quantitative vergleichende Analysen
- TV 3.2: Technik: Visualisierung, Gridifizierung, Modellierung der Vernetzung



Geisteswissenschaftliche Teilvorhaben

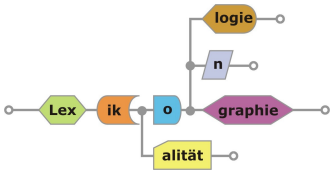
Linguistik



INSTITUT FÜR
DEUTSCHE SPRACHE

Institut für deutsche Sprache in Mannheim (IDS)

- TV 1: 'Erstellung der Standardlemmaliste'
- IDS betreut die Entwicklung einer Lemmaliste der nhd. Standardsprache
- als Netzstruktur (Verweisstruktur) organisiert
- erlaubt den semasiologischen Zugriff auf sämtliche Varietäten
- Ausgangspunkt: DeReWo-Korpus und DeReWo-Liste
- Weiterbearbeitung



Auszug aus der DeReWo-Liste mit einfacher XML-Struktur: Zähler, Lemma, Häufigkeitsgruppe

V:\glmetall\DATEN\DeReWo\derewo.xml

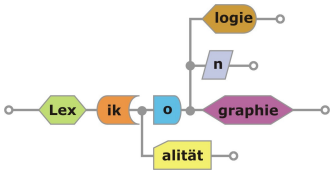
```
<entry id="29970" lemma="Karriereleiter" group="17" />
<entry id="29971" lemma="detoniert" group="17" />
<entry id="29972" lemma="Passagierzahl" group="17" />
<entry id="29973" lemma="Götterdämmerung" group="17" />
<entry id="29974" lemma="Exzellenz" group="17" />
<entry id="29975" lemma="Elegie" group="17" />
<entry id="29976" lemma="hervorholen" group="17" />
<entry id="29977" lemma="Wasserglas" group="17" />
<entry id="29978" lemma="Ursachenforschung" group="17" />
<entry id="29979" lemma="Sozialgeschichte" group="17" />
<entry id="29980" lemma="Reinigungsmittel" group="17" />
<entry id="29981" lemma="Mitbürgerin" group="17" />
<entry id="29982" lemma="Fontäne" group="17" />
<entry id="29983" lemma="Doktortitel" group="17" />
<entry id="29984" lemma="Zentralorgan" group="17" />
<entry id="29985" lemma="Vereinskasse" group="17" />
<entry id="29986" lemma="Okkupation" group="17" />
<entry id="29987" lemma="Monopoly" group="17" />
<entry id="29988" lemma="Leerstelle" group="17" />
<entry id="29989" lemma="Konfirmand" group="17" />
<entry id="29990" lemma="Kerzenlicht" group="17" />
<entry id="29991" lemma="Filmreihe" group="17" />
<entry id="29992" lemma="Einsender" group="17" />
<entry id="29993" lemma="Bastille" group="17" />
<entry id="29994" lemma="Badesaison" group="17" />
<entry id="29995" lemma="ausspionieren" group="17" />
<entry id="29996" lemma="appetitlich" group="17" />
<entry id="29997" lemma="allerwenigste" group="17" />
<entry id="29998" lemma="Verkehrsstrom" group="17" />
<entry id="29999" lemma="Philologe" group="17" />
</derewo>
```

Fertig

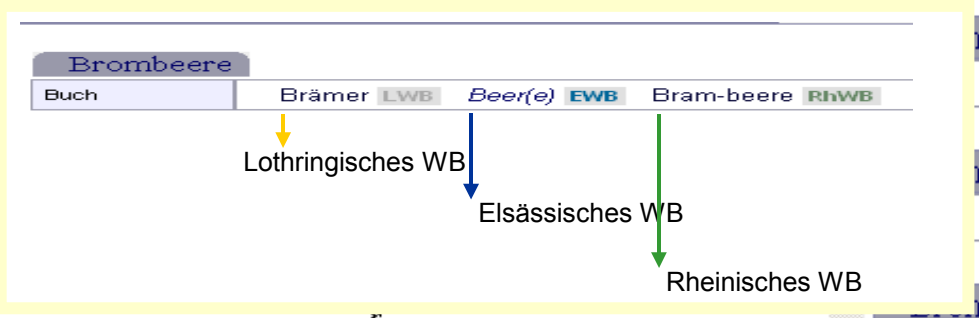
Verbundkoordinator: Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften an der Universität Trier

- TV 2: 'Erstellung von klassifizierten Varietätenlemmalisten', semasiologische Vernetzung, Erkundungsmodul onomasiologische Vernetzung
- Datenbasis: Digitale Wörterbücher, Wörterbuchnetz
- Exzerpiert werden Lemmaansätze und ggf. Lemmavarianten sowie soweit als möglich entsprechende Symptomwerte (Raum, Zeit, Textsorte etc.)
- Varietätenlemmalisten werden mit der Basislemmaliste nach semasiologischen Kriterien verknüpft (Methodiken in TV 3)
- Netzstruktur entsteht

- Erkundungsmodul: onomasiologische Vernetzungstrukturen



Brombeere f.: wie schd., zur Verbr. vgl. K. 68, s. auch Karte Brombeere in DWA X, *Brambeer (brambēr)* [verbr. im pfälz.-elsäss. Grenzgebiet PS-Lembg KL-Frankst NW-Weidth Hardbg Dürkh Wachh LU-Alsh/Gr Limbghf LA-Altd GH-Hayna Rh'zab], *Brombeer (brom-)* [KU-Schönbg HB-Brenschb PS-Geisbg KL-Hütschhs Nd'mohr O'sulzb Rockhs schmales Gebiet zwischen Kaisl u. Frankth FR-Albch Flomh Eppst Raum zwischen Spey u. Germh (neben *Brami*



[PS-O'simt Gal-Sap Buch-A' Schindhd Erfw Geisbg FR-H Germh GH-Hagb Don Franzt [verbr. im Raum zwischen St. Horschb RO-Schneebghf Mü Gal-Moosbg Bruckenth Ottenhs Rosenbg Fehlb], *Brembier* [Kus KU-Etschbg Theisbgstg HB-Alth IB-Bebh Pirmas PS-Erlbn Don-Alexanderhs Tschest St. Andreas Buch-Tereblestie], *Brimbeer* [PS-Merzalb BZ-Annw Rambg], *Blambeer* [verbr. im Raum südl. Lu'haf (vgl. BERTRAM 141)], *Blombeer* [GH-Sondhs], *Blembeer* [GH-Knitth Bellh], *Brame* [»Zweibrücker Ggd., Westrich vielfach« (WILDE 34) LU-Friesh], *Breme (brę̄mə, selten brēmə)* [verbr. nördl. WPf NPf nördl. VPf BZ-Billh verbr. Don Gal Buch], *Bremet (brę̄mətd)* [KL-Mölschb W'lein], *Bremde (brēmdə)* [im Gebiet nordwestl. KU-Wolfst], *Brime (brīmə)* [RO-Hallgt O'mosch Nd'mosch Als], *Brumel (brumə̄l)* [PS-Rodalb BZ-Apphf Kling Mühlhf Germh GH-Hagb], *Bremel (brę̄məl)* [KB-Stauf FR-Carlsbg A'lein NW-Ungst in einem Gebiet nordwestl. FR-Grünstdt Don-Ulmbach Liebling Gal-Zboiska], *Bremel* [KB-Rams], *Blame (blāmə)* [FR-Flomh NW-Hardbg Wachh], *Bleme (blę̄mə)* [im Gebiet um FR-Grünstdt (zusammen mit *Blame*)], *Flambee* [LU-Schauh Assh Iggh NW-Forst Deidh Ruppbg LA-Queichh], *Flame* [im Raum Frankth (zusammen mit *Brombeer*)], *Flome* [FR-Bockh LU-Opp], *Flembee* [GH-Zeisk Hördt], *BrÄfl* [GH-Nd'lustdt O'lustdt] (dasselbe Wort auch in dem etwa auf gleicher Höhe östlich des Rheins gelegenen Rußheim). Zs. Heckenbrombeere. Syn.: Brummels-, Dorn-,

Brombeere
Buch: Brämer LothrWB *Beer(e)* ElsWB; Bram-beere RheinWB

nbeeren-kirbe
Brambeerkirb RheinWB

nbeeren-saft
Saft ElsWB

nbeeren-wein
Buch: Brambeerwein RheinWB

Promenade
Buch: Promenade RheinWB

promenieren
Buch: promenieren RheinWB

pro nobis
Buch: pronobissen RheinWB

Pröpeler
Buch: brebeln RheinWB

Propeller
Buch: Propeller RheinWB

Bezugssysteme: Wörterbücher

mhd. Wbb (BMZ und Lexer)

brâmber *stn. brombeere. sumerl.* 40,70. 56,77. 57,53.
1, 258. daz hulfe in niht ein brâmbere *Mone altd. schausp.* 3,44

brâm-ber *stn.* (I. 104[a]) *brombeere* MONE *schausp.* EILH.
swarz geverwet als ein zîtic brâmber TROJ. 32743; bromber HPT.

hist. nhd. Wb. (Grimm)

BROMBEERE, *f. rubum,* brambeere, *schwankt in* branbeere,
Bd. 2,
branbire, bramber, braunbeere, brobeere, braubeer, brommer. DASY.
brombeer *morum rubi.*

Pfälz. Wb

Brombeere f.: wie schd., zur Verbr. vgl. K. 68, s. auch Karte
in DWA X, *Brambeer (brambēr)* [verbr. im pfälz.-elsäss. Gren

Rhein. Wb

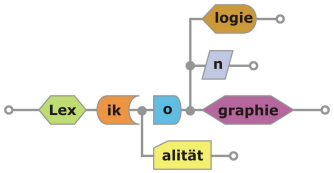
Bram-beere Formen. mit erhaltenem *-beere* u. *brâm-* > *brom-:*
brombęs, Pl. *-z* Klevld [*brom lbęr n* Rees-Brünen, *brøm lbęs*

Els. Wb

Brambeer, Brombeer, Brombere, Bromer(e), Bromert, Brämer,
Bromter, Blombeer, Blomber, Blomere [Prämpêr Ndröd. Wingen Lohr;
Prompér Co.; Prümpêr Ensish. Urbis Su. Mutzig Str. K. Z. Betschd.;
Prümpərə Pfetterhsn.; Prömərə Roppenzw. Hirzfn.; Prómər Co. Rchw.

Lothr. Wb

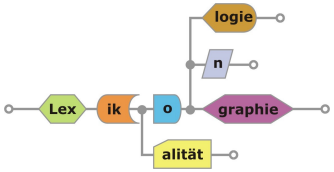
Brämer [brémər, bréïmər *Si.*] *m. Brombeerstrauch.* – els. 2, 189
Brämer; mhd. brâmber.



standardisiertes Bezugssystem: Metalemmaliste

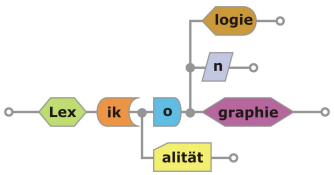
Metalemma (nhd. Standard, öffentlicher Sprachgebrauch)

- gemeinsames Drittes
- verbindet unterschiedliche Wortparadigmen der Varietäten
- Schnittstelle für externe Informationssysteme (z.B. Ontologien)
- Kernbestand vorhanden
  - Auf der Basis des IDS-Referenzkorpus Liste der 30.000 häufigste Wörter als Grundlage für die Meta-Lemmaliste
  - Teil-Lemmalisten (als Teil-Bezugssystem) zu spezifischen Bereichen, z.B. Sprachstufe Mittelhochdeutsch oder westmitteldeutsche Dialekte
- Varietätenlemmata müssen mit Metalemmata in Bezug gesetzt werden
- Bildung von Pseudo-Metalemmata, z.B. für ausgestorbene Wörter



naturwissenschaftliches Teilvorhaben

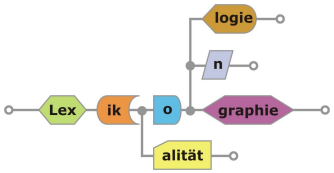
Bio-Informatik

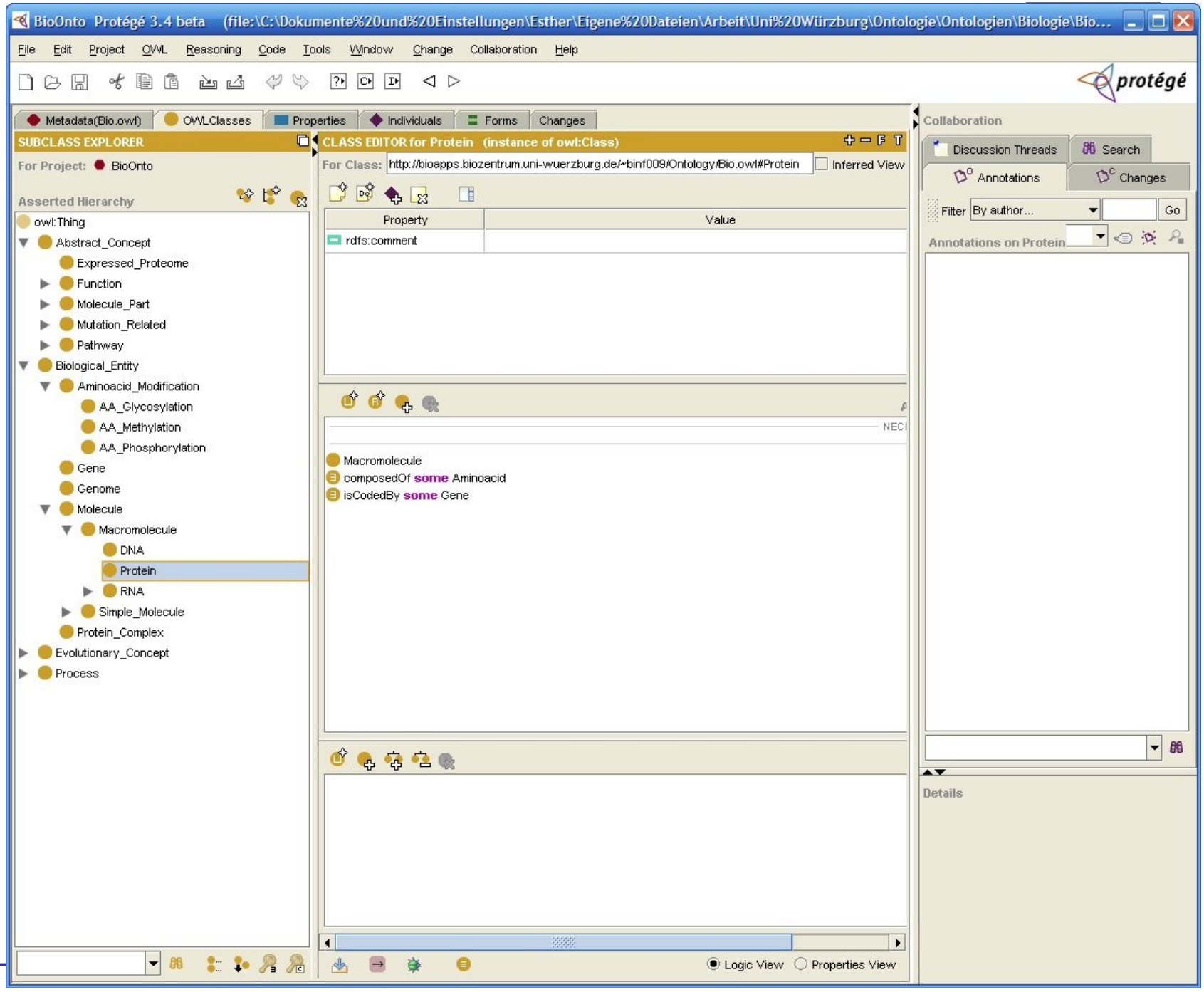


BIOZENTRUM
UNIVERSITÄT
WÜRZBURG

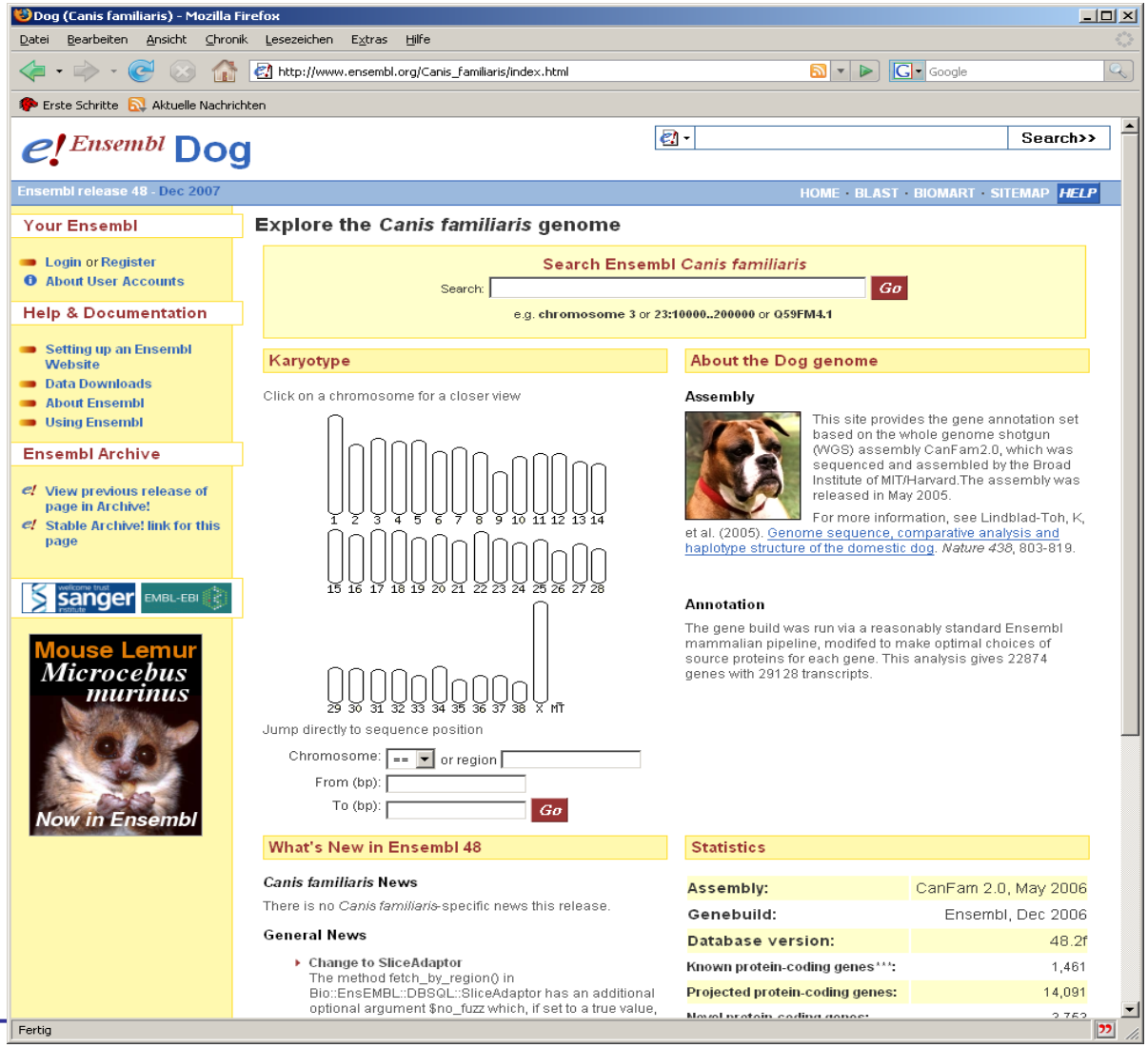
Theodor Boveri-Institut für Biowissenschaften/Bioinformatik,
Universität Würzburg

- TV 3.1: Grammatik der Varianz in Genomen und Sprache, Quantitative vergleichende Analysen
- Identifikation analoger Prozesse
- quantitativer Vergleich dieser Prozesse
- Erstellung von Ontologien der beteiligten Entitäten sowie der Varianz erzeugenden Mechanismen für Genome und Sprachen
- Ontoverse-Software (www.ontoverse.org)
- Modellierung von Objekten und ihren Beziehungen in Sprache (Wort, Morphem, Buchstabe) und Genomen (Protein, Domäne, Aminosäure)
- Integration ausgewählter Varianz erzeugender Mechanismen (Wort- / Genfusion, Wort- / Genverlust, horizontaler Transfer)
- Mapping der Ontologien für Genome und Sprache



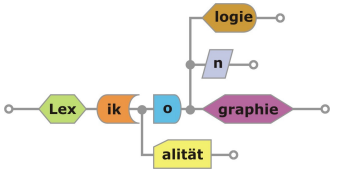


Datenbasis Biologie
Genomische Datenbank



Frei verfügbar

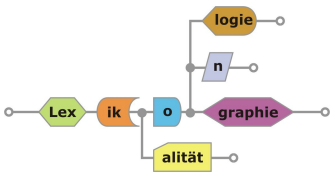
www.ensembl.org



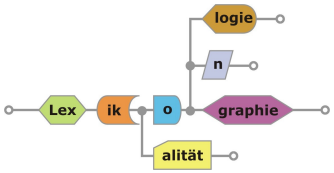
Bundesministerium
für Bildung
und Forschung

integratives Teilvorhaben

Informatik

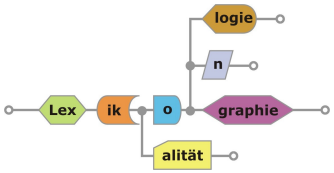


- TV 3.2: 'Technik: Visualisierung, Gridifizierung, Modellierung der Vernetzung'
- Organisation der effizienten Verarbeitung und Speicherung der Daten
- Zusammenführung der Konzepte von Genomik und Sprachwissenschaft
- adäquate Präsentation der Daten
- Visualisierungen von Beziehungen und Vernetzungen
- Integration der Ergebnisse in TextGrid



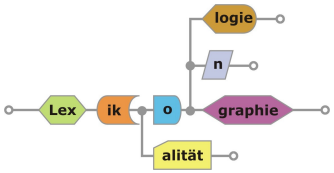
vorläufiger Entwurf
Seipel / Schneiker / Borek

```
<metalemma id="1" value="gut" pos="adjective" source="ids">
  <lemma corpus="ids" corpus_id="234" value="gut" lang="nhd"
    frequency_class="4" />
  <lemma corpus="lexer" corpus_id="5424" value="guot"
   lang="mhd" >
    <references corpus="bmz" corpus_id="122" />
    <example>
      bistu übel ode guot lw. 27.
    </example>
  </lemma>
  <lemma corpus="bmz" corpus_id="122" value="guot"
   lang="mhd">
    <references corpus="grimm" corpus_id="534" />
  </lemma>
    <lemma corpus="grimm" corpus_id="534" value="guot"
   lang="mhd" />
</metalemma>
```

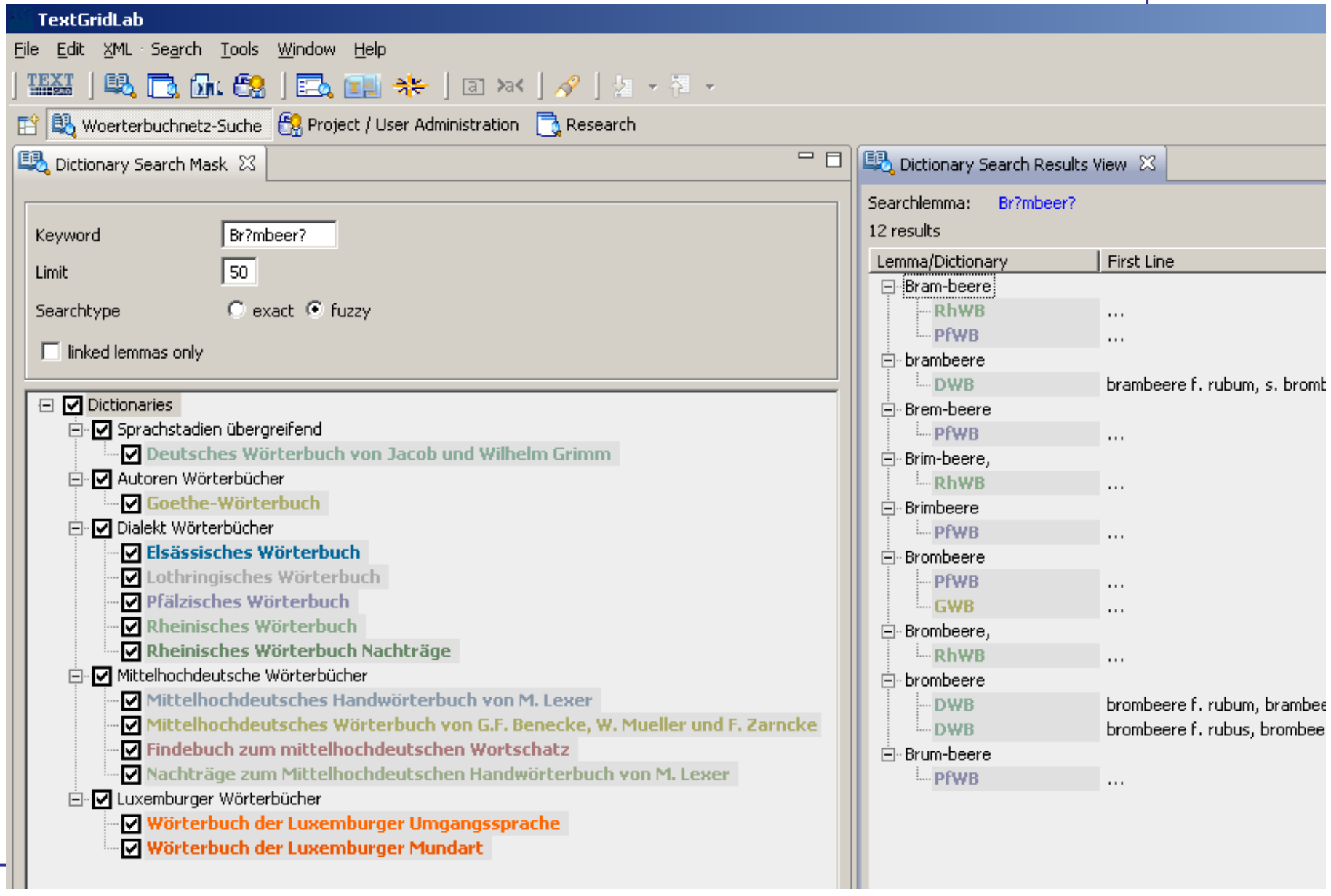


weitere Kooperationen

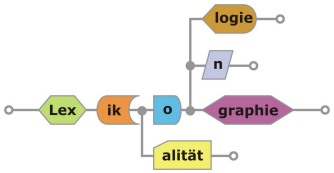
- Wörterbuchbuchkanzleien
- Projekt regionalsprache.de (REDE): Forschungsplattform zu den modernen Regionalsprachen des Deutschen (Forschungszentrum Deutscher Sprachatlas, Marburg)
- TextGrid: Modulare Plattform für verteilte und kooperative wissenschaftliche Textdatenverarbeitung - ein Community-Grid für die Geisteswissenschaften
- ...



Kooperationen: z.B. TextGrid

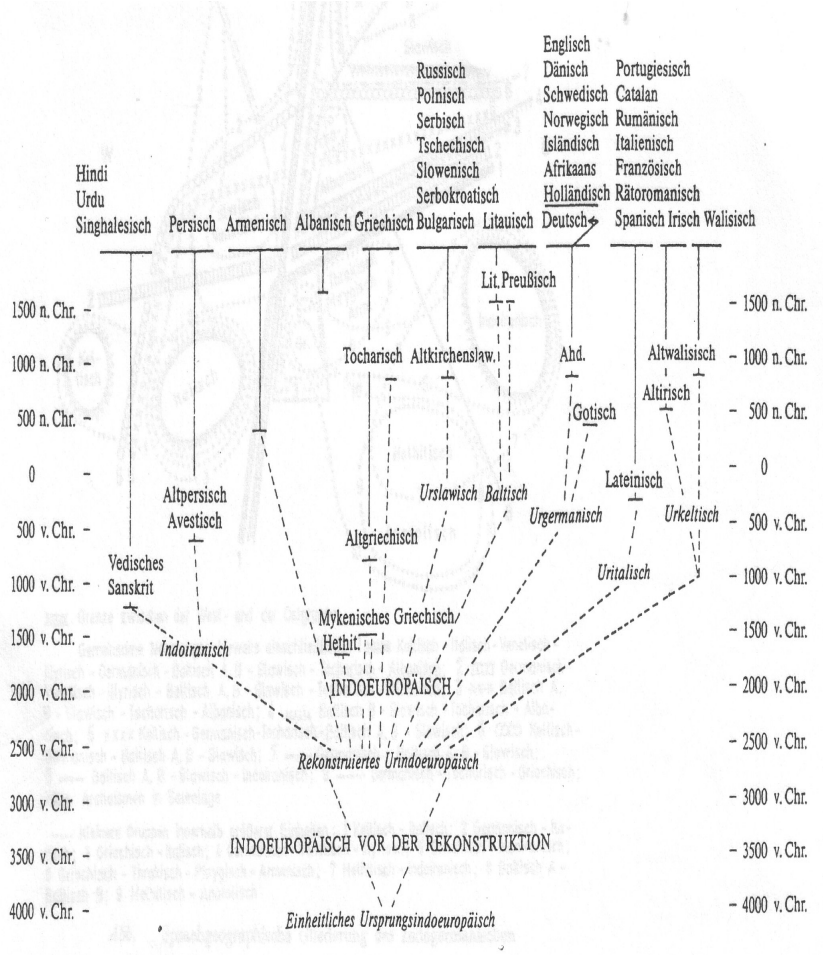


Identifikation analoger Mechanismen

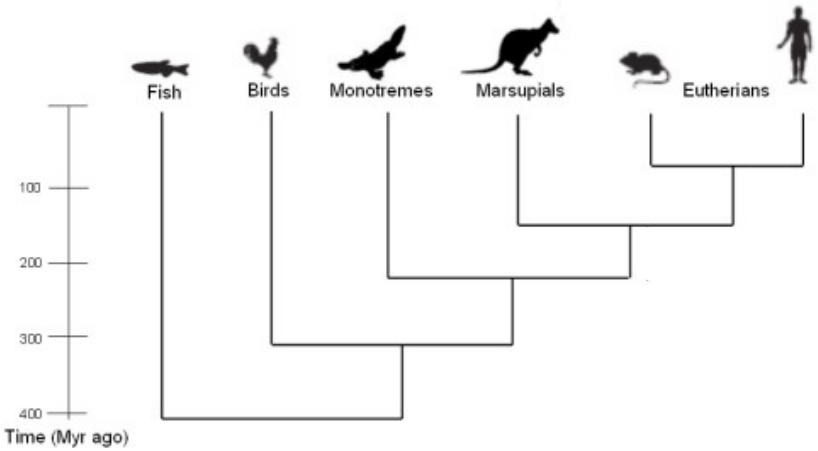


Evolution in der Linguistik und der Biologie

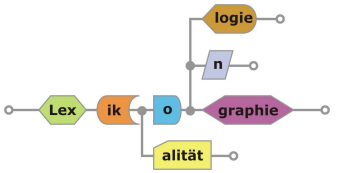
Die Entwicklung der idg. Sprachen



Die Entwicklung der Vertebraten (Wirbeltiere)

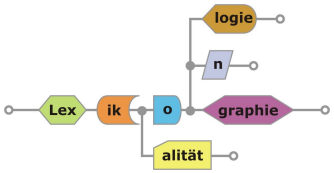


- Unterschied der Zeiträume!



aus: R.E. Keller, Die deutsche Sprache und ihre historische Entwicklung. Hamburg 1995, S. 32.

regionale und historische Varianz

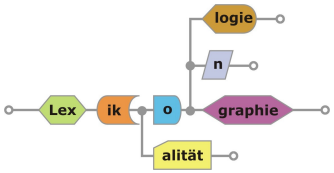


Varianz in der Sprache am Beispiel ‚Brombeere‘

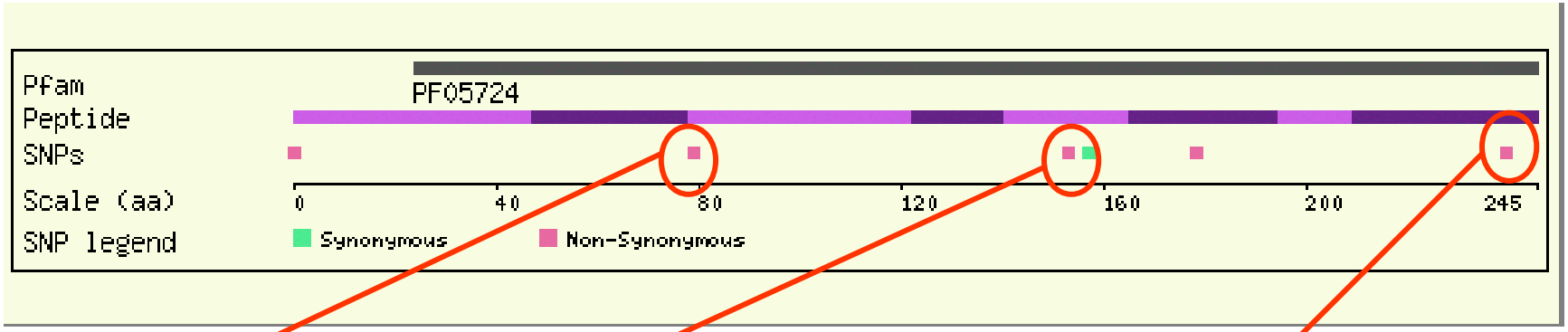
Parameter: semasiologisch (ausdrucksseitig)

idg.	**bher-*	hervorstehen, etwas Spitzes (z.B.
idg.	**bha-*	glänzen Dornen)
ahd.	*brāma, prāma*	Dornbusch
mhd.	*brâme*	Dornbusch
ahd.	*brām-beri*	Brombeere
mhd.	*brâm-ber*	Brombeere
nhd.	*Brom-beere*	

Brombes	Kleve
Brommelte	Düsseldorf
Bromele	Bonn
Brambel	Saarlouis
Bromere	Schlettstadt



Thiopurine S-methyltransferase loss-of-function Mutationen

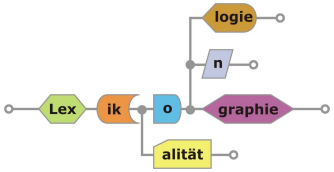


TMPT*2 A->P

TMPT*3A A->T
Häufigste Variante in Amerikanisch / Kaukasisch

TMPT*3C Y->C
Häufigste Variante in Afro-Amerikanern

Regional verteilte Varianten beeinflussen Dosierung von Medikamenten



Varianz in der Sprache am Beispiel ‚Brombeere‘

Parameter: onomasiologisch (inhaltsseitig, semantisch)

nhd. *Brom-beere*

Dornbeer

Kratzbeer

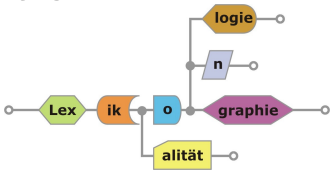
Schmärzbeer, Schmäsbeer, Schmelzbeer

Heckenbeer, Heckbes, Heckebraimcher

Tranbeer, Bachbeer,

Schwa(r)zbeer, Schmoazbel, Schwosbat, Woschbert

Perdsbeer, Fuchsbeer, Katzentapen, Katzedoube



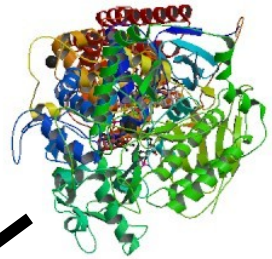
Neue Einheit – gleiche Funktion

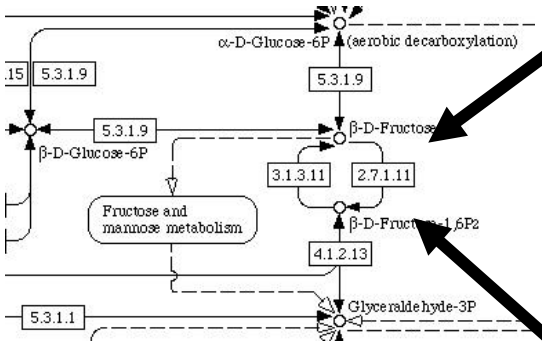
Onomasiologische Varianz

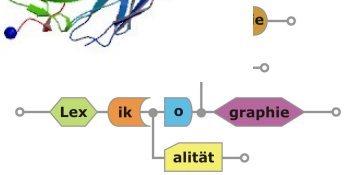
Brombeere

Fuchsbeere

Nicht orthologe Gen-Ersetzung

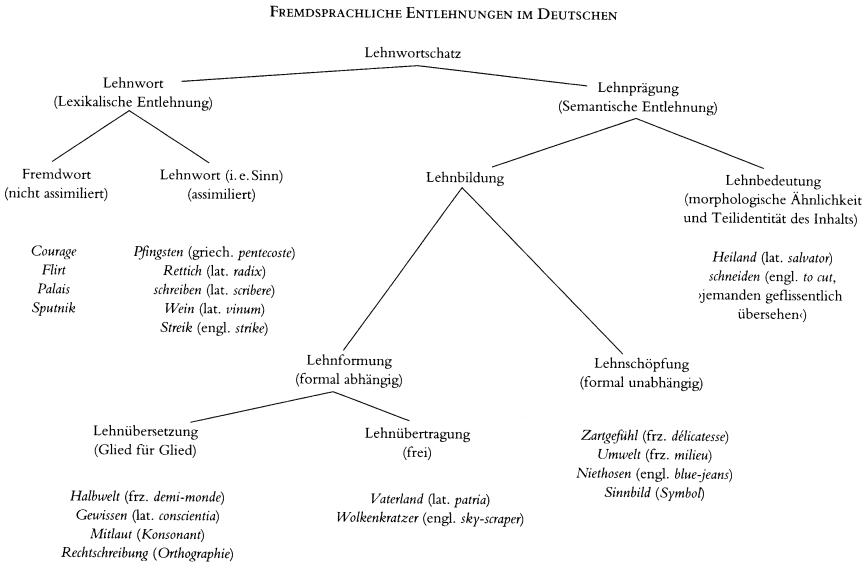




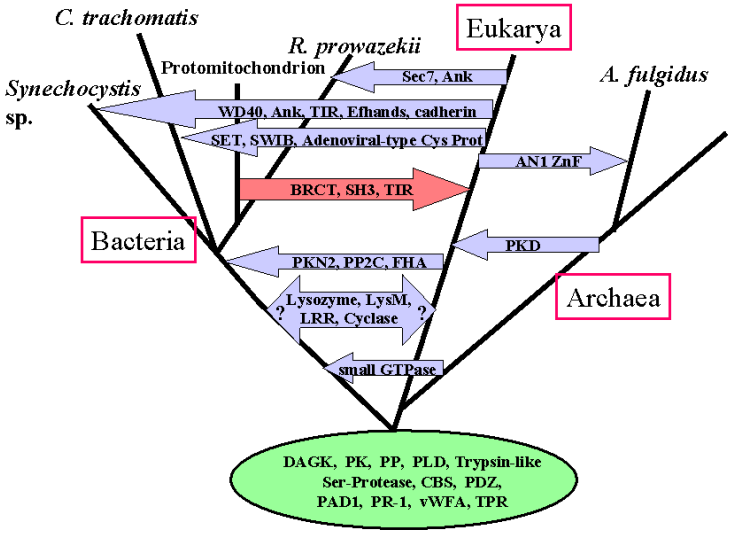


Horizontaler Transfer vs. Entlehnung

Entlehnungsmechanismen



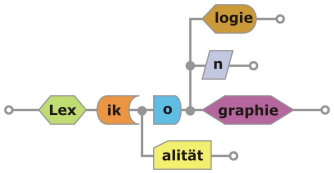
Horizontaler Gentransfer



Strukturen:

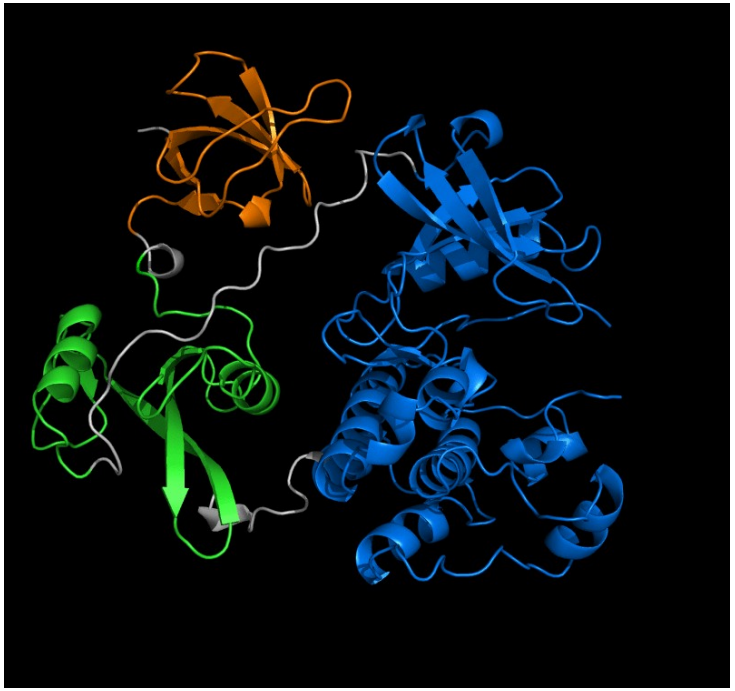
Morpheme

Domänen



Domänen in Proteinen

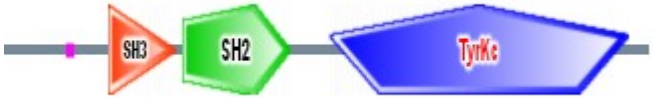
Struktur

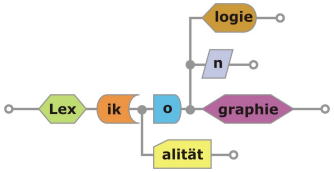


Sequenz

```
MEPFLRKRLAFLSFFWNKIWPAGGEPVSDGSWSPDAEFPAEPGSPPAPARLFRALSDFRA
RCAGELSVSRGDWLFALQEDGEYLFARRLSSPFCTGLVPITCVAKTSPEPLSDQPWYFGD
ISRTQAQQLLLSAANAPGAFLVRPSESRRGDYSLSVRAQAKVCHYRISMAEDGGLYLQKG
QCFSSLEELLSYYKANWKLLQNPLLQPCMPQKSQEQDQWERPRSEFVLRRKLGQGFFGEV
WEGLWLGSVPVAVKVIKSADMKLDDLAQEIRTLKSLRHERLIRLHAVCSTGEPVYIVTEL
MRKGSLQAYLGGPGGRTLSLPLLLSFACQVAEGMGYLECRRIVHRDLAARNVLVGDNLAC
KVADFGLARLLKDDVYSPSSSSKIPVKWTAPEAANYCIFSQKSDVWSFGVFLYEVFTYGQ
CPYEGMSNHETLQQVMQGYRLPRPPTCPAEVYVLMLECWKGSPEERPAFSVLQEKLGDIS
RCFYP
```

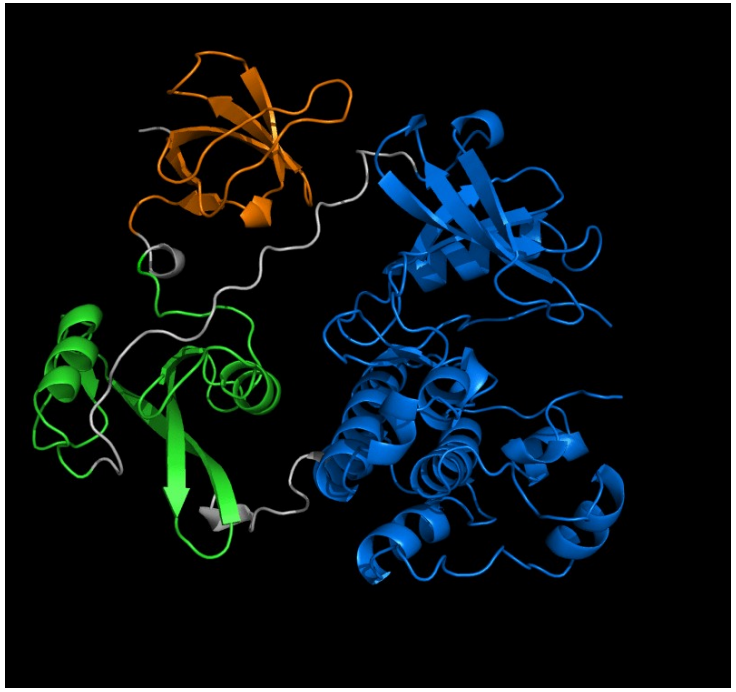
Domänen





Domänen in Proteinen

Struktur

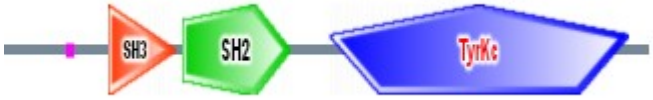


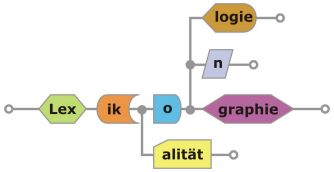
Sequenz

MEPFLRKRLAFLSFFWNKIWPAGGEPVSDGSWSPDAEFPAEPGSPPAPARLFRALSDFRA
RCAGELSVSRGDWLFALQEDGEYLFARRLSSPFCTGLVPITCVAKTSPEPLSDQPWYFGD
ISRTQAQQLLLSAANAPGAFLVRPSESRRGDYSLSVRAQAKVCHYRISMAEDGGLYLQKG
QCFSSLEELLSYYKANWKLLQNPLLQPCMPQKSQEQDQWERPRSEFVLRRKLGQGFFGEV
WEGLWLGSVPVAVKVIKSADMKLDDLAQEIRTLKSLRHERLIRLHAVCSTGEPVYIVTEL
MRKGSLQAYLGGPGGRTLSLPLLLSFACQVAEGMGYLECRRIVHRDLAARNVLVGDNLAC
KVADFGLARLLKDDVYSPSSSSKIPVKWTAPEAANYCIFSQKSDVWSFGVFLYEVFTYGQ
CPYEGMSNHETLQQVMQGYRLPRPPTCPAEVYVLMLECWKGSPEERPAFSVLQEKLGDIS
RCFYP

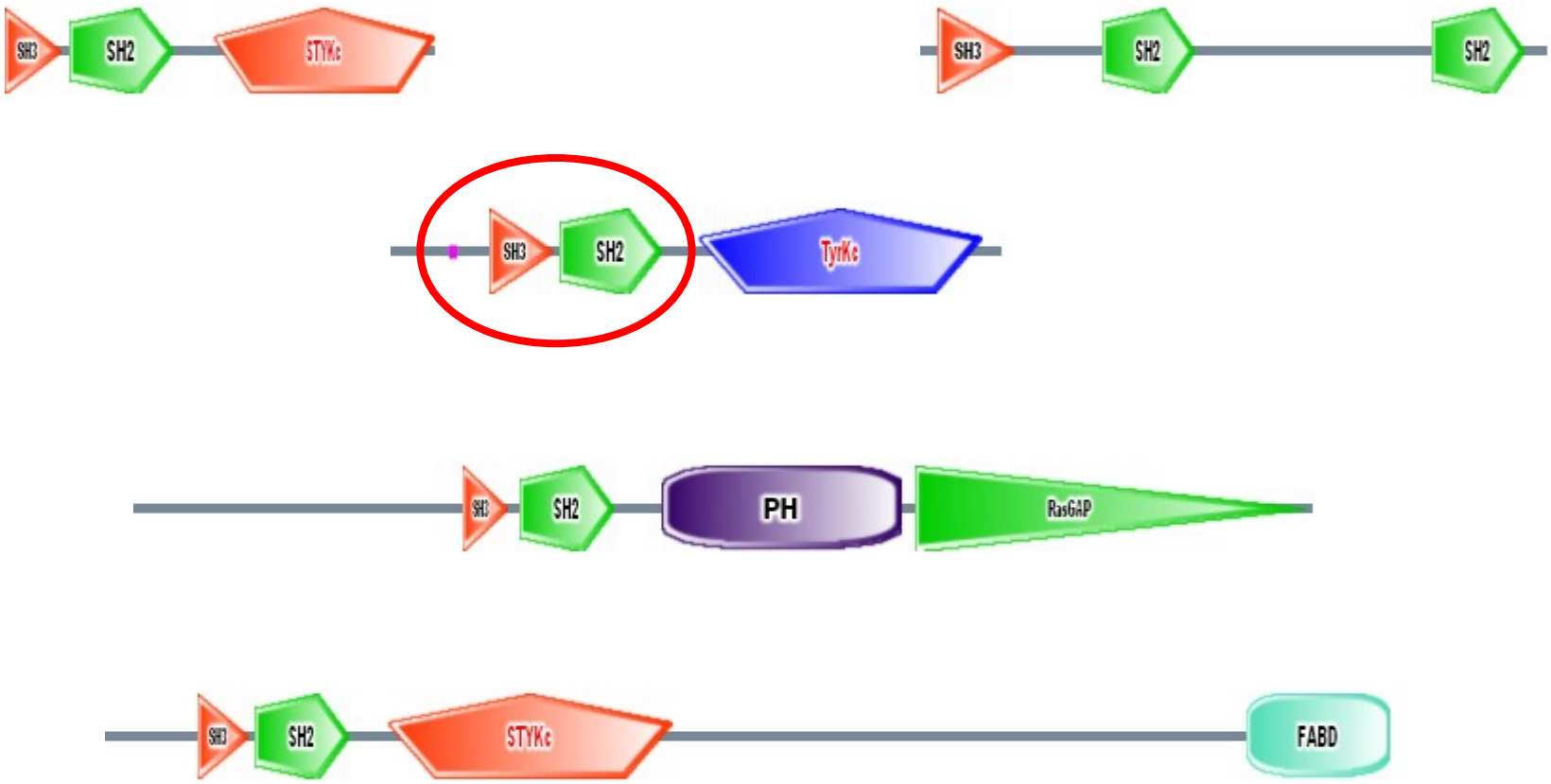
Hidden Markov Modelle

Domänen

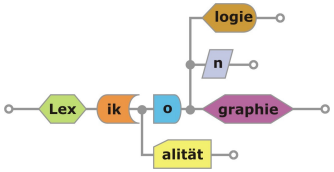




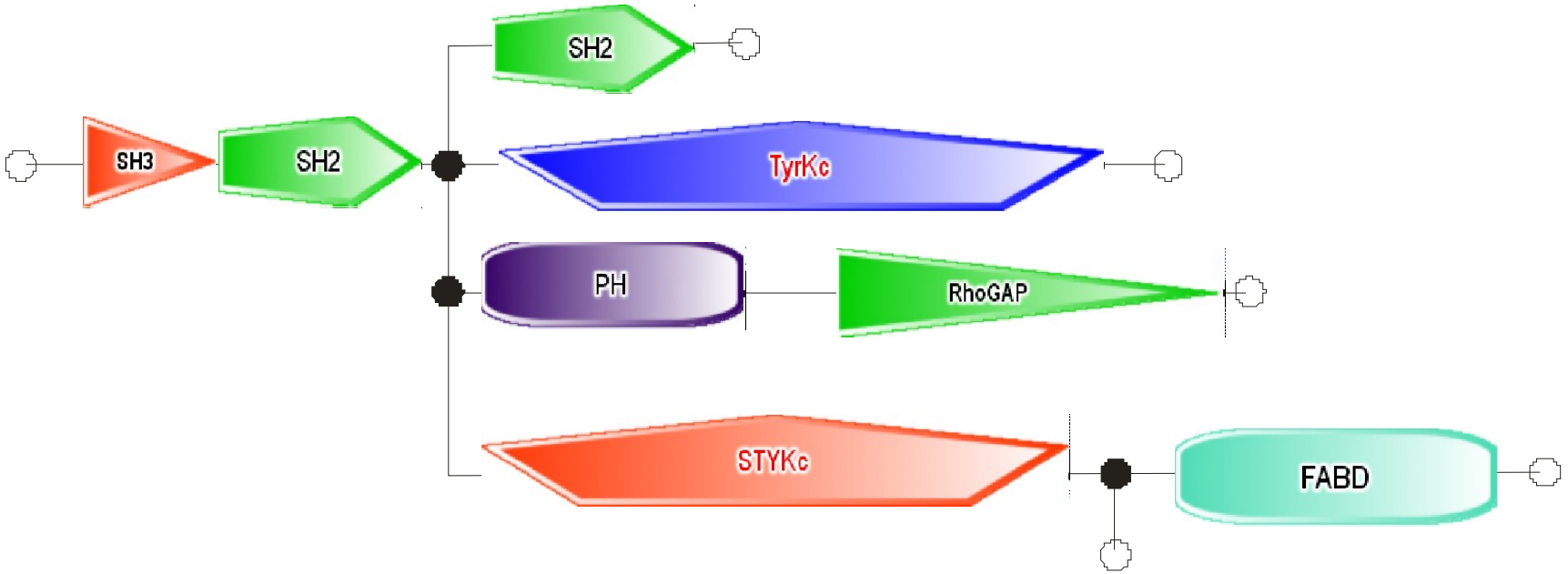
Lexikon von Proteindomänen
Weitere Einträge mit Proteinstamm?



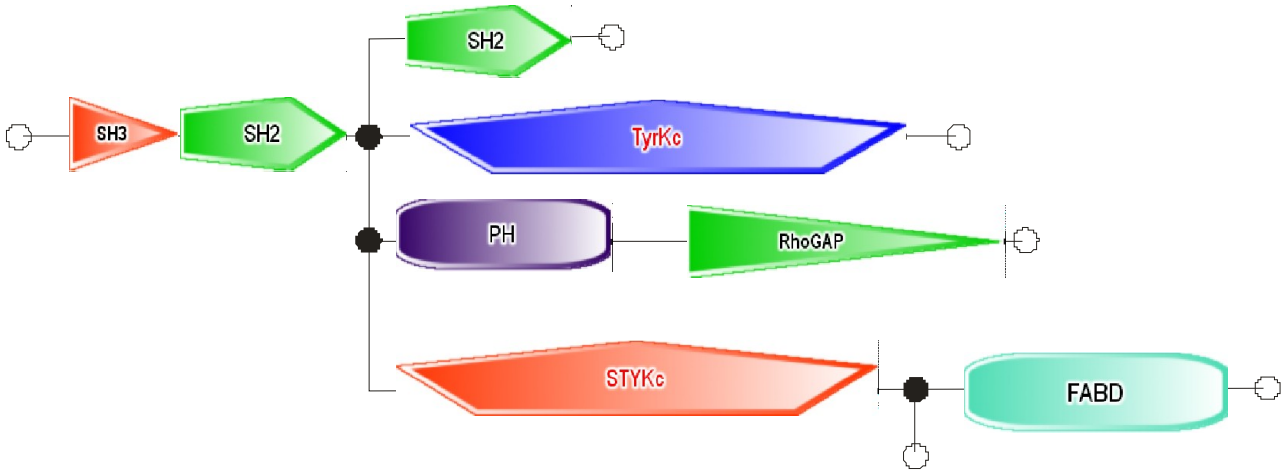
http://smart.embl-heidelberg.de

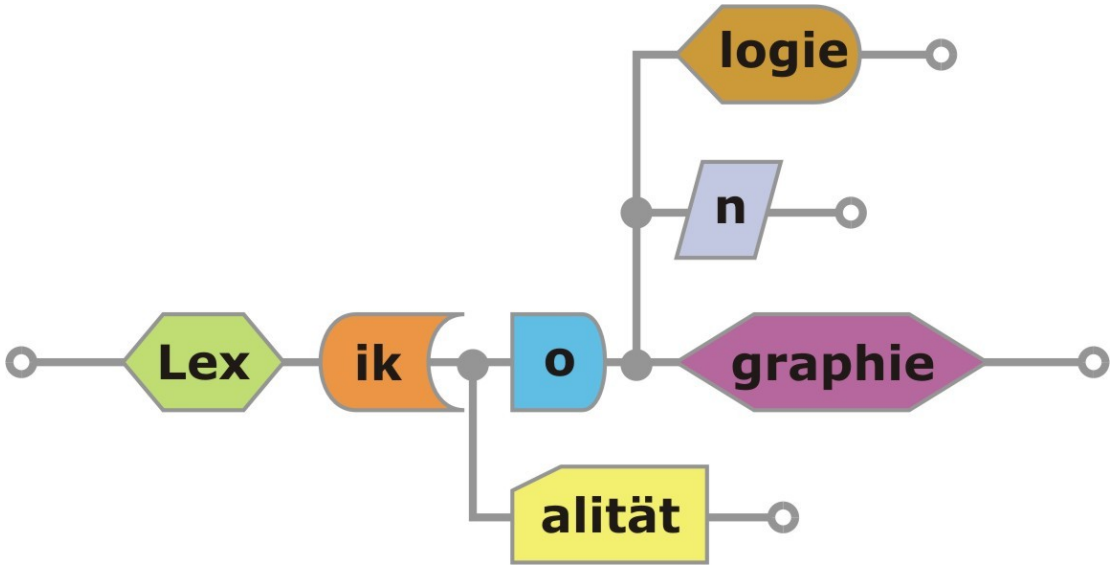


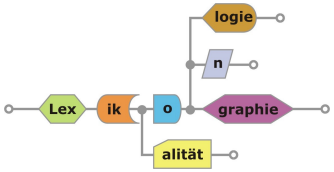
Domänen in Proteinen



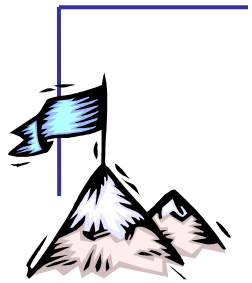
Domänen und Morpheme



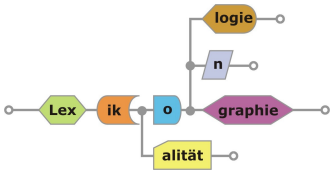




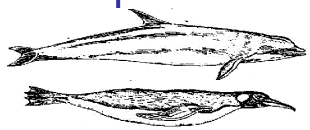
Ziel



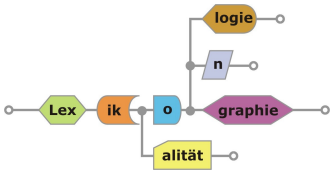
- Vergleich biologischer und sprachlicher Prozesse
- Systematische Identifikation von analogen Konzepten und Mechanismen
- Empirisch-quantitative Evaluierung dieser Prozesse



Identifikation analoger Konzepte und Mechanismen

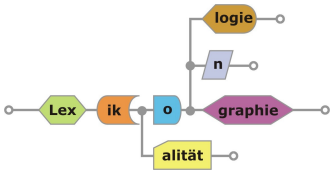


- Ontologie Varianz erzeugender Mechanismen
  - Sprache
  - Genomik
- Mapping der Ontologien
- Identifikation von
  - analogen Prozessen
  - transferierbaren Methoden



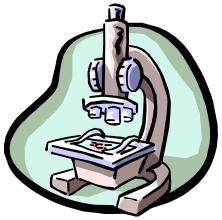
Informatik als integrierende Infrastruktur

- Beide Felder nutzen informatische Methoden, z.B.
  - Statistik (z.B. HMM),
  - Data Mining/Text Mining (Clustering),
  - regelbasierte Verfahren,
  - Information Retrieval,
  - Grammatiken/FSA
- Datenbanken
- Datenintegration



Empirisch-quantitative Evaluierung

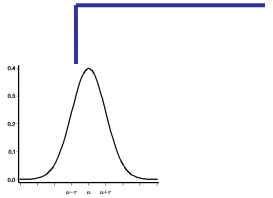
Gehorchen die Prozesse
gleichen empirischen Gesetzen?



Benötigt werden:

- Infrastruktur
- Daten

Datengrundlage

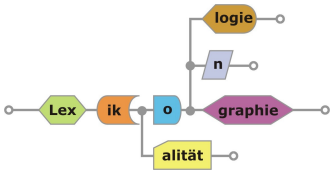


- Biologie
  - Verbundpartner beteiligt an der Erarbeitung der benötigten Datenbasis
  - Genomprojekte (Mensch, Maus)

→ Bezugssystem vorhanden

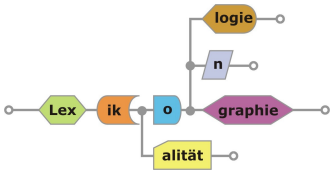
- Sprachwissenschaft
  - Verbundpartner beteiligt an der Erarbeitung der benötigten Datenbasis
  - digitale Wörterbücher (Lemmalisten)
  - Dialektdatenbanken

→ Bezugssystem fehlt: Meta-Lemmaliste

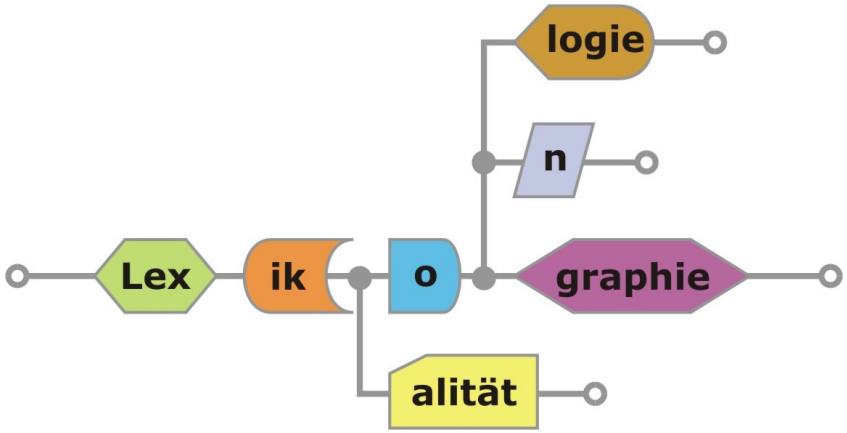


Anwendungs- und Einsatzmöglichkeiten für linguistische Seite des Projekts

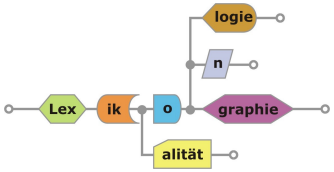
- neue und präzisere Verfahren der Informationsgewinnung
- Varianz kann z.B. für phonologisch-graphonematische, morphologische und lexikalische Analysen des Wortschatzes über eine Meta-Lemmaliste ausgeglichen werden, ohne dass dadurch das sprachliche Basismaterial verfälscht wird
- Recherche über Metalemmaliste ermöglicht Suche auf varianten Korpora über standardisierten Zugang
- Verbesserung von Rechercheergebnissen
- Lemmatisierung varianter Texte
- ...



Herzlichen Dank für Ihre Aufmerksamkeit!



Dr. Andrea Rapp, Kompetenzzentrum Trier



Biologisch-linguistisches 'Wörterbuch'

Aminosäuren	20	Phoneme	13-75 (Deutsch ca. 39)
Domänen	5.000	Basismorpheme	5.000-10.000 (Selbständig)
		Wortbildungs-morphem/Flexionsmorphem	700-1.000 (Mobil)
Proteine/Proteinkomplexe	60.000-10.0000	Wörter	ca. 40.000-100.000 +
Gene	20.000	Wortbildungsbaupläne (Basismorph.+Flexionsm.)