

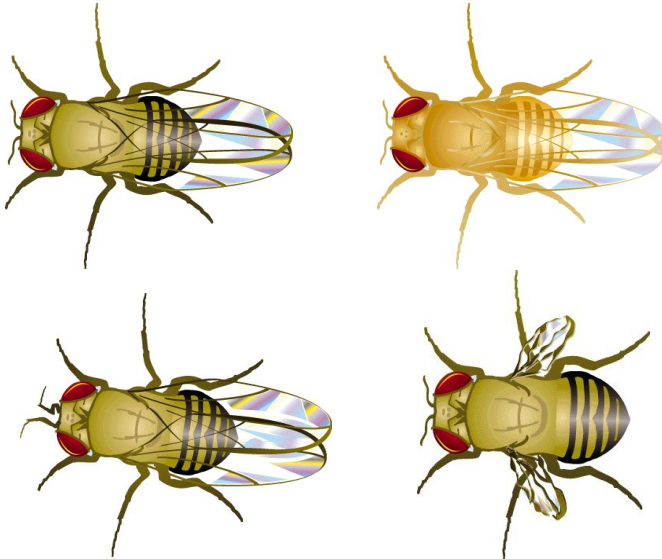
Wechselwirkungen zwischen linguistischen und bioinformatischen Verfahren, Methoden und Algorithmen

Modellierung und Abbildung von Varianz in Sprache und Genomen

Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren in den
Geisteswissenschaften
Universität Trier

18. November 2009

Varianz I - Mutation bei *Drosophila Melanogaster*



Quelle: www.exploratorium.edu/exhibits/mutant_flies/mutant_flies.html



Varianz II - Varietätenlemmata

vliege

fliege

Fliege

vleuge

fliuga

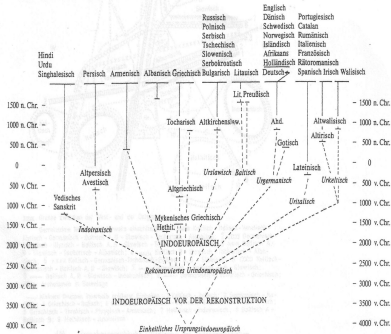
vliuge

Mögliche Analogien

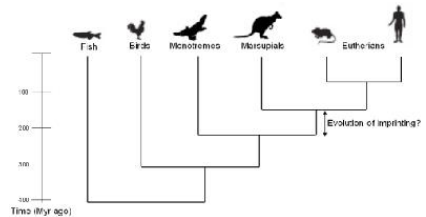
- Varietätenlemmata
- Komposita
- Neuschöpfungen mit vorhandenen (Flexions-)bausteinen
- → *Mutation* = „Abwandlung der Stammformen durch Lautwechsel“
Bußmann, H. *Lexikon der Sprachwissenschaft*. Kröner: Stuttgart, 1983.
- ...

Ursprung der Analogien

Entwicklung der idg. Sprachen



Entwicklung der Vertebraten



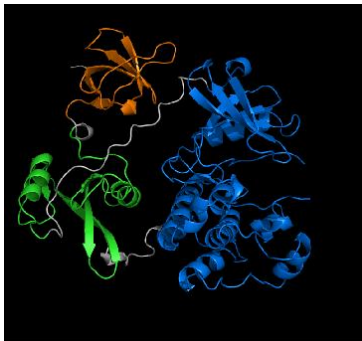
aus: R.E. Keller, Die deutsche Sprache und ihre historische Entwicklung. Hamburg 1995, S. 32.

Was ist eine Mutation?

„... a typing error made in copying a text“

Grassé, Pierre-Paul. *Evolution of Living Organisms*. Academic Press: NY, 1977.

Domänen in Proteinen



MEPFLRKRLAFLSFFWNKIWPAGGEPVSDGSWSPDAEFPAEPGSPAPARLFRALSDFRA
RCAGELSVSRGDMFLAQEDGEYLFARRLSSPFCITGLVPITCVAKTSPEPLSDQWYFGD
ISRTOAQQLLLSAANAPGAFLVRPSESRRGDYLSVRAQAKVCHYRISMAEDGGLYLQKG
QCFSSLEELLSSYKANWKLQNPLLQPCMPQKSQEQDOWERPRSEFVLRRLGQGGFFGEV
WEGMLGSVPVAVKVIKSADMKLDDLAQEIRTLKSLRHERLIRLHAVCSTGEPVYIVTEL
MRKGSLOAYLGGPGGRTLSPDLLSFACQVAEGMGYLECRRIVHRDLAARNVLVGDNLAC
KYADFGRLARLLKDDVYSPSSSSKIPVKWTAPEAANYCIFSQKSDVWSFGVFLYEVFTYGQ
CPYEGMSNHETLQQVMQGYRLPRPPTCPAEVYYLMLECWKGSPEERPAFVLQEKLGDIS
RCFYF



„Tippfehler“ im Mutanten *wingless*

279 AACGGGAATTCGCTCGATACATTCAATATACCAAAACCATTCGCAAAACAACAACAACCTCGAAGGGAAGTATCTATCAT
EcoRV
358 ACCCGGTGTGTCAAGTGTGAGAGTGTGTGTGCGCTCGAAACAGATAAACCGCATACGAATAATGGATATACAGTATATCT
1MetAsp11eSerTyr11eP
437 TGTTCATCTGCCCTGATGCCCTGTGTCAGCGCGCGCAGCTCTCAGCCAACTCGAGGGCAACAGAAATCGGAAGGGG
7pHeVal11eCysLeuMet1AlaLeuCysSerGlyGlySerSerLeuSerGlnValGlyGlyLysGlnLysSerGlyYArgp
516 CCGGGGCTCCATGTGTGGGCATTCGCAAGGTGGCGCAACCAACAACATTACGCCCATCATGTATCATGGACCCAGGG
33pYArgGlySerMetTrpTrpGly11eAlaLysValGlyGlyuProAanAen11eThrPro11Met1TyrMetAspProAla
Neri
595 ATCCACTCTACGTTGAGAGGAAACAGCGACGCGCTGCTCAGGCAACAATCCGGTGTATCTGGGAGCCCTGGTCAAGGGG
60P11eHeSerThrLeuArgArgLysGlnArgArgLeuValArgAspAanProGlyValLeuGlyAlaLeuValLysGlyY
674 CCAACTTGGCCATTTCAGAGTCGCCAACCAAGTTCAGAAATCGCGCGCTGGAACTGCTCGACGACCAAACTCTCGAGGG
81AlaenLeuAla11eSerGlyuCysGlnH1eS11eArgArgAanArgTrpAenSerTrpAanPheSerArgGly
A
wg
IL14
Sar
753 CAAAAATCTATTTCGCAAAATCGTTGATCGAGGCTGCGGAGAGACAGGCTTCATTACCGAATCAGCCAGCGCGCGGT
112pLysAanLeuPheGlyLysLeuValAspArgGlyCysArgGlyuThrSerPhe11eTyrAla11eThrSerAlaAlaVal
BglII
832 ACCCATCTCGATTGCCAGGGGCTGCAGCTGAAGGAACGATAGAGTCTCGACCTGCGACTACAGCCACCGTGGAGATCTC
139pThrHiSer11eAlaArgAlaCysSerGlyGlyThr11eGlySerThrCysAspTyrThrHiS11eGlySerArgP
911 CACAAGCGAACCAACAGCGGGCAGTGTGGCGGCGCTGCGGGATTGGAGTGGGCGGCTCGCCTCGACAACTCGGATT
165pArgAlaAlaAanHiGlyAlaGlySerValAlaGlyValArgAspTrpGlyuTrpGlyGlyCysSerAspAan11eGlyPhe
EcoRI
990 CGGGTTCAGTCTCTCCGGGAATTCGTCGATACCGCGAGAGCGGCTGCAATCTCGCGGAGAGATGAATCTGCAAC
191pGlyPheLysPheSerArgGlyuPheValAspThrGlyGlyuArgGlyArgAanLeuArgGlyLysMetAanLeuHiAan
A
wg
DN7
1069 AACGAGGCGCGCTCGAGGCCACGTCGCAAGCGGAGATGCGACAGGAGTGCAAATGCCATGGCATGTCCGGATCGTGTACAG
218pAanGlyuAlaGlyArgAlaHiValGlyAlaGlyuAlaArgGlnGlyuLysCysHiGlyMet1SerGlySerCysThrV
A
Bsp
1148 TGAAGACTCTGCTGGATGGCGCTGGCCAACTCTCGGTGTATTGGCGAACATCTGAAGGCGCGCTTCGATGGAGCCACCGG
244pLysThrCysTrpMetArgLeuAlaAanPheArgMet11eGlyAspAanLeuLysAlaArgGlyAspGlyYArgp
EcoRI
1277 CGTGCAGGTGACCAACAGCTCTCGGGGCCACCAACAGCTCTCGGCCCACTTAGTTCGAATGCGCGGCTCGAATTCCTGT
270pValGlnValThrAanSerLeuArgAlaThrAanAlaLeuAlaProValSerProAanAlaMetGlySerAanSerVal

1 Das Projekt

- Vorstellung des Projekts

2 Die Metalemmaliste

- Einordnung in das Projekt
- Datengrundlage
- Konzeption der Metalemmaliste
- Zuordnungsverfahren

3 Sprachwissenschaftliche Ontologie

4 Morphemzerlegung

„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.

GEFÖRDERT VOM



Bundesministerium
für Bildung
und Forschung

Vorstellung der Verbundpartner

Verbundkoordinator

- Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften

Prof. Dr. Claudine Moulin

Verbundpartner

- Institut für Deutsche Sprache Mannheim (IDS)

Prof. Dr. Ludwig Eichinger

- Kompetenzzentrum für EDV-Philologie an der Universität Würzburg

Prof. Dr. Werner Wegstein

- Biozentrum Universität Würzburg

Prof. Dr. Jörg Schultz

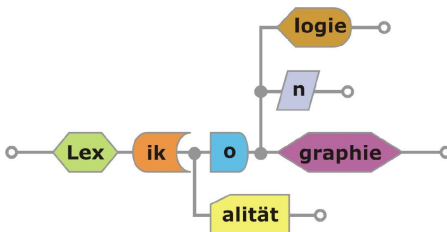
- Lehrstuhl für Informatik I Informationsstrukturen und wissensbasierte Systeme

Prof. Dr. Dietmar Seipel

Projekttitel

Wechselwirkungen zwischen linguistischen und bioinformatischen
Verfahren, Methoden und Algorithmen

Modellierung und Abbildung von Varianz in Sprache und Genomen



Gemeinsame Forschungsfrage

- prinzipielle strukturelle Gemeinsamkeiten
 - zwischen Genomcode und sprachlichem Code
 - biologischer Entwicklung und Sprachentwicklung
- zentrale gemeinsame Kennzeichen
 - 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

TV1

- Basislemmaliste der neuhochdeutschen Standardsprache

IDS Mannheim

TV2

- Klassifizierte Varietätenlemmalisten, semasiologische Vernetzung;
Erkundungsmodul onomasiologische Vernetzung

Kompetenzzentrum, Uni Trier

TV3.1

- „Grammatik der Varianz in Genomen und Sprache“, Quantitative
vergleichende Analysen

Bioinformatik, Uni Würzburg

TV3.2

- Technik: Visualisierung, Gridifizierung, Modellierung der Vernetzung


Informatik, Uni Würzburg

Standardisiertes Bezugssystem

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 Metalemmaliste
 - 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

Datenbanken der Disziplinen



Home - Fruitfly 2007-11-13

Login | Register | BLAST/BLAT | BioMart | Docs & FAQs | Mirrors

About this species

Description

Genome Statistics

Assembly and Genebuild

Top 40 InterPro hits

Top 500 InterPro hits

What's New

Sample entry points

Karyotype

Location (2L:21850001-2)

Gene (cul-2)

Transcript (cul-2-RA)

Variants (not available)

Configure this page

Manage your data

Export data

Bookmark this page

Search Ensembl Fruitfly

Search for:

Go

Description

Assembly and Genebuild

Fruitfly (*Drosophila melanogaster*)

Assembly and Annotation

The data displayed on the Ensembl Fruitfly site is a compendium of data from different sources, including [BDGP](#), [FlyBase](#) and [DGRP](#). The genomic sequence is based on BDGP assembly release 5, and annotations displayed in Ensembl are imported from FlyBase release 5.4 (dated 01 November 2007). No additional gene build has been carried out, however [transcript](#) and [protein](#) features have been compiled through the Ensembl pipeline.

We have included *Drosophila* into the Ensembl system to allow people to access the Fly genome through the Ensembl user interface (both for visualisation and data mining) and to provide cross-species integration through our comparative genomics resources (such as homologous gene links and family pages).

The canonical data for *Drosophila* is managed at [FlyBase](#) and we are working in partnership with them to provide feedback on their resources.

Ensembl release 56 - Sept 2009 © [HTC](#) | [ER](#)

[About Ensembl](#) | [Contact Us](#) | [Help](#)

[Permanent link](#) | [View in history](#)

Ensembl www.ensembl.org

Deutsches Wörterbuch von Jacob Grimm und Wilhelm Grimm. 16 Bde. Im 32 Teilbandn. Leipzig: S. Hirzel 1854-1890. – Quellenverzeichnis 1871.

Benutzersuche | Vorworte | Wörterbücher im Wörterbuchnetz | Der digitale Grimm auf CD-ROM

Stichwort suchen

FRUCHTFLIEGE, f. - FRUCHTHECKE, f. (Band 4, Spalten 272 - 273) ▶

Gliederung | Vernetzung | Suchen

A FRUCHTFLIEGE, f. eine Fliege, deren Larve in Früchten oder samen lebt, carposmia.

B FRUCHTFOLGE, f. die aufeinanderfolge und ordnung des baues der einzelnen fruchtarten auf demselben grundstücke oder felde von jahr zu jahr. man kann z. b. dasselbe feld nicht jedes jahr mit rogen bepflanzen, sondern es muss die fruchtfolge eine andre sein. vgl. fruchtumlauf, fruchtwechsel.

C FRUCHTGARTEN, m. ein obstgarten, dann überhaupt ein mit obestäumen, essbare beeren tragenden sträucher, gemüse u. dgl. beplanzter garten, zum unterschiede von einem lust- oder ziergarten.

FRUCHTGÜLTE, f. Synon. Frucht-gulte [Pflanzw](#)

FRUCHTHECKE, f. Berechnet (ausgegrüht) GELÄNDERBAUM, m. [DWB](#)

FRUCHTHOLZ, n. Berechnet (ausgegrüht) FRUCHTZWIG, m. [DWB](#)

FRUCHTLAND, n. Berechnet (ausgegrüht) GEHÖLZREICH [DWB](#)

FRUCHTRUTHE, f. Synon. Frucht-rute [Pflanzw](#)

Deutsches Wörterbuch www.dwb.uni-trier.de



Luise Borek

Wechselwirkungen

Lemmalisten/Wörterbuchdaten

■ Wörterbücher

- Deutsches Wörterbuch von Jacob Grimm und Wilhelm Grimm
- Mittelhochdeutsches Wörterbuch von Benecke, Müller und Zarncke
- Mittelhochdeutsches Handwörterbuch von Matthias Lexer
- Findebuch zum mittelhochdeutschen Wortschatz
- Pfälzisches Wörterbuch
- Rheinisches Wörterbuch
- Wörterbuch der elsässischen Mundarten
- Wörterbuch der deutsch-lothringischen Mundarten
- Goethe-Wörterbuch
- Wörterbuch der deutschen Gegenwartssprache

■ Korpusdaten

Basislemmaliste

-<derewo>

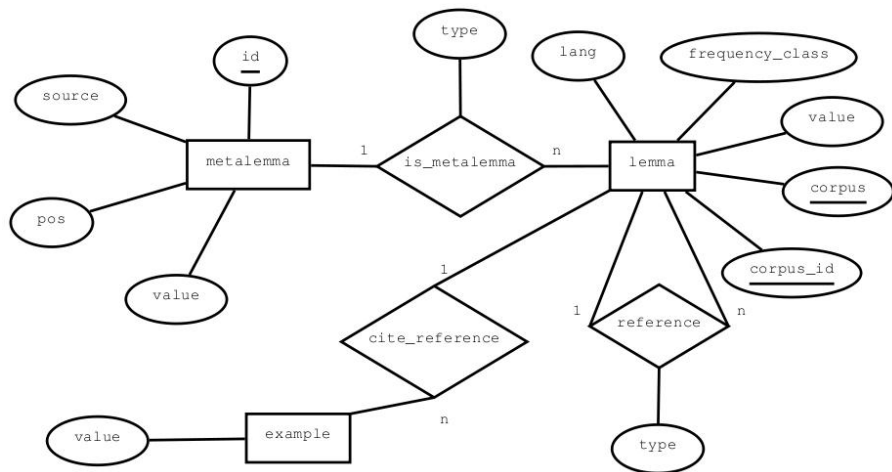
```
<entry id="0" lemma="d-" group="0"/>
<entry id="1" lemma="in" group="2"/>
<entry id="2" lemma="und" group="2"/>
<entry id="3" lemma="sein" group="2"/>
<entry id="4" lemma="ein" group="3"/>
<entry id="5" lemma="zu" group="3"/>
<entry id="6" lemma="werden" group="3"/>
<entry id="7" lemma="von" group="3"/>
<entry id="8" lemma="haben" group="3"/>
<entry id="9" lemma="mit" group="4"/>
<entry id="10" lemma="er/sie/es/sie" group="4"/>
```

...

```
<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>

ER-Diagramm



```
<metalemma id="1" value="Fliege" pos="nn" source="bll">
  <lemma corpus="bll" value="Fliege" lang="nhd">
    <lemma corpus="dwb" value="fliege">
      <lemma corpus="gwb" value="Fliege">
        <lemma corpus="rhwb" value="Fliege">
          <lemma corpus="pfbw" value="Fliege">
            <lemma corpus="bmz" value="vliege">
              <lemma corpus="lexer" value="vliege">
                <lemma corpus="lexer" value="vleuge">
                  <lemma corpus="lexer" value="fliuga">
                    <lemma corpus="findebuch" value="vliege">
                      <lemma corpus="findebuch" value="vliuge">
</metalemma>
```

Komplexes Metalemma

```
<metalemma id="128" value="August" pos="nn"
  source="bll">
  <lemma corpus="bll" corpus_id="414"
    value="August" lang="nhd"/>
  <lemma corpus="gwb" value="August"/>
  <lemma corpus="rhwn" value="August"/>
  <lemma corpus="rhwb" value="August"/>
  <lemma corpus="rhwb" value="Augst"/>
  <lemma corpus="pfb" value="August"/>
  <lemma corpus="dwb" value="augst"/>
  <lemma corpus="elswb" value="Augst"/>
  <lemma corpus="lexer" value="ougst"/>
  <lemma corpus="lexer" value="ougeste"/>
  <lemma corpus="lexer" value="ougst"/>
  <lemma corpus="lexer" value="ougste"/>
  <lemma corpus="lexer" value="augustö"/>
  <lemma corpus="lexer" value="ochste"/>
  <lemma corpus="lexer" value="ogest"/>
  <lemma corpus="lexer" value="ougwest"/>
```

```
<lemma corpus="lexer" value="oust"/>
<lemma corpus="lexer" value="ouste"/>
<lemma corpus="lexer" value="ouwest"/>
<lemma corpus="lexer" value="ouwestinne"/>
<lemma corpus="lexer" value="owest"/>
<lemma corpus="lexer" value="owist"/>
<lemma corpus="lexer" value="öugestinne"/>
<lemma corpus="nachtrlexer" value="ougest"/>
<lemma corpus="findebuch" value="ougest"/>
<lemma corpus="findebuch" value="augst"/>
<lemma corpus="findebuch" value="oust"/>
<lemma corpus="findebuch" value="ouwest"/>
<lemma corpus="bmz" value="ougeste"/>
<lemma corpus="bmz" value="ougste"/>
<lemma corpus="bmz" value="oust"/>
<lemma corpus="bmz" value="ouwest"/>
<lemma corpus="bmz" value="owest"/>
</metalemma>
```

Zuordnungen

Zuordnungsverfahren

- Vernetzung diachroner und diatopischer Varianten des Deutschen
- Verankerungspunkt ist dabei ein nhd. Basislemma
- Generelle Möglichkeiten der Zuordnung:
 - linguistisches Expertenwissen
 - Verfahren des maschinellen Lernens/Statistik
- Hilfestellung und Qualitätssicherung
 - Zusätzliche Informationen aus den Wörterbüchern (z. B. part of speech)
 - Verweise in den Wörterbüchern

Beispiel

- Durch den Einsatz von Lautgesetzen auf ältere Sprachstufen lässt sich die neuhochdeutsche Form (re)konstruieren

- Beispiel

	Mhd.	Nhd.
Neuochhochdeutsche Diphthongierung	⟨î⟩ →	⟨ei, ai⟩
	⟨û⟩ →	⟨au⟩
	⟨iu⟩ →	⟨eu, äu⟩
	bew ^î sen →	bewe ^{ei} sen
	t ^û chen →	ta ^{au} chen
	sch ⁱ une →	Sche ^{eu} ne

Selbständiges Lernen: Sequence Alignment

Idee

- Motivation aus der Bioinformatik
- Anordnung von Strings so, dass *score* maximiert wird
- Grundidee ist, dass ein String aus dem anderen durch „Mutation“ hervorgegangen ist
- Beispiel:

u	n	g	e	a	c	h	t	e	t
u	n	g	e	a	-	h	t	e	t
- *Score* bewertet mittels Parameter: Alignierung übereinstimmender Buchstabensymbole, Alignierung unterschiedlicher Symbole, Lücken (-)
- Bestimmen der Parameter aus linguistischen Datensätzen

Übersicht

	In Basislemmaliste	Nicht in Basislemmaliste
In den Wörterbüchern	verknüpfbar	Pseudometalemma notwendig; Erweiterung der Basislemmaliste erforderlich
Nicht in den Wörterbüchern	z. B. Entlehnungen jüngeren Ursprungs (<i>Football</i>)	–

Vorteile und Anwendungen

Vorteile und Anwendungen

- Erleichterte Suche durch Erschließung von Varietätenlemmata; flexibler Einstieg über das Metalemma
- Erhalt der originalen Wörterbuch-Strukturen bei Mehrwert durch Vernetzung
- Ergänzung durch Onomasiologie
- Andockstelle für weitere Informationssysteme

„Ontology Alignment“

Ausrichtung der Ontologie

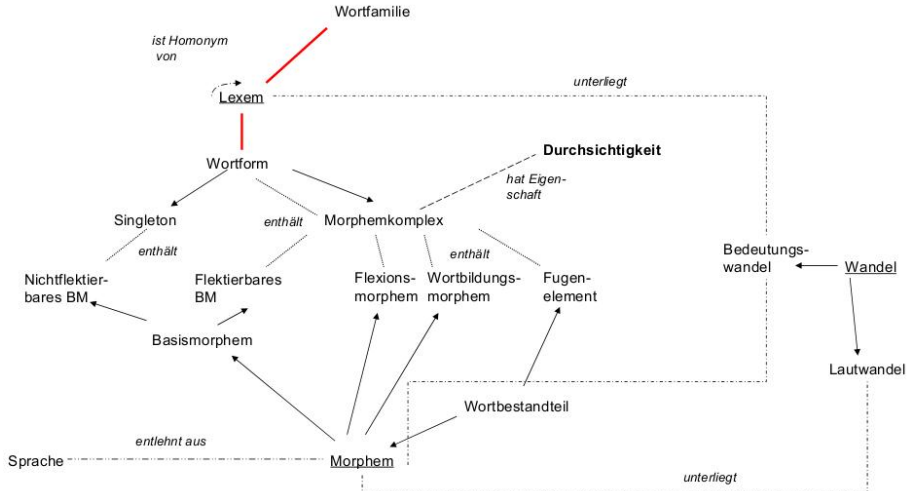
- Werkzeug um Daten aus verschiedenen Quellen und Bereichen aufeinander abzubilden
- Methodologisches Aufzeigen von Analogien zwischen den Disziplinen
- Die sprachwissenschaftliche Ontologie wird unabhängig von der existenten biowissenschaftlichen Ontologie aufgebaut
- Der Schwerpunkt liegt auf den zu untersuchenden Begriffen und Relationen
- Auch für andere Forschungsfragen verwendbar und beliebig erweiterbar

„Ontology Alignment“

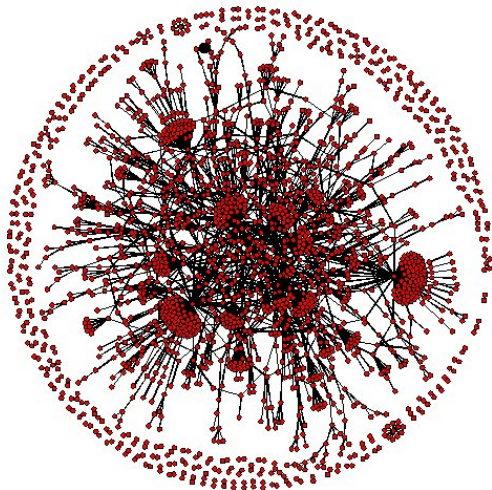
Ziele des „Ontology Alignment“

- Linguistik: Organisation und Aufbereitung von Daten; Auffinden neuer Muster?
- Bioinformatik/Biowissenschaft: Rückschlüsse auf nicht vorhandenes historisches Material

Sprachwissenschaftliche Ontologie



Zum Morphemnetzwerk des DWB



Morpheme Network aus dem DWB mit *R* (Daniela Keller, WÜ)

Viele Nachbarn

Niedriger Clusterkoeffizient

Grad	Morphem	ClusCoef
2829	ung	0,01299
2336	ver	0,01659
2164	un	0,01280
1075	auf	0,01972
733	ei	0,01385
680	in	0,00918
638	heit	0,01411
615	viel	0,01835

Hoher Clusterkoeffizient

Grad	Morphem	ClusCoef
559	haupt	0,10377
545	stueck	0,10272
500	meister	0,10384
484	recht	0,10055
467	mensch	0,10871
453	schlag	0,10021
439	hof	0,10487
437	sonn	0,10265

Wenige Nachbarn

Höchster Clusterkoeffizient. . .

Grad	Morphem	ClusCoef
2	abendlich	1
2	action	1
2	bastei	1
2	behelmt	1
3	aufklaer	1
3	aufleg	1
4	erpress	1
4	faehnrich	1

Herzlichen Dank!

Anhang

Übrigens...

Kreative Gennamen für *Drosophila*-Mutanten

- *tinman* - Embryos ohne Herz
- *cheap date* - Reagiert sehr stark auf Alkohol
- *maggie* - Entwicklung bleibt stehen (in Anspielung auf Maggie Simpson)
- *gypsy* - Das mutierte Gen ist ein mobiles genetisches Element
- *swiss cheese* - Viele Löcher im Gehirn
- *Indy/I'm not dead yet* - Doppelte Lebensdauer
- *Krüppel*
- *Windbeutel*
- *Nudel*

Weitere Domänen mit Proteingruppe

