

Data and Metadata of Language Resources as Linked Data on the Web: Examples

Christian Chiarcos

chiarcos@uni-frankfurt.de



Linked Data for Linguistics



- Representation and modelling
- Structural interoperability
- Integrating distributed resources
- Conceptual interoperability
- Dynamic Import
- Ecosystem, infrastructure and community

Representation and modelling



- Different linguistic subcommunities have developed representation standards, e.g.,
 - LMF: Lexical Markup Framework (Francopoulo et al. 2009)
 - lexical-semantic resources
 - GrAF: Graph Annotation Framework (Ide and Suderman 2007)
 - annotated corpora
 - based on labeled directed acyclic graphs (feature structures)
- RDF data model: labeled directed graphs
 - Uniform formalism for *different* resource types
 - Sublanguages (e.g., OWL) to define specific vocabularies

Structural interoperability



- With different language resources represented in RDF, we can combine both sources of information freely
 - cross-resource queries with SPARQL
 - Given a corpus with WordNet sense annotations
 - “Retrieve all sentences that describe locations”
 - i.e., sentences containing a token annotated with a WordNet sense that is a hyponym of “location”
- Difficult to realize with GrAF or LMF

Integrating distributed resources



- SPARQL supports nested subqueries to run on different repositories

```
SELECT ?token {
  service <http://wordnet.rkbexplorer.com/sparql> {
    rkbWN:synset-land-noun-2
      wn20:containsWordSense ?sense .
      ?sense rdfs:label ?synonym .
  }
  ?token powla:hasString ?synonym .
}
```

- No physical integration of resources in a single data base required
 - Easy to link to centralized repositories of reference terminology, etc.

Conceptual interoperability



- Resources should specify which vocabulary (e.g., for annotation) they use and how it is defined
 - By reference to community-maintained terminology repositories, e.g.,
 - GOLD (Farrar and Langendoen 2010)
 - ISOcat (Windhouwer and Wright @ LDL-2012)
 - OLiA (Chiarcos 2008)
 - Can be used, e.g., for disambiguation
 - e.g., *land* as a noun, but not as a verb

Dynamic import



- Linking resources with URIs
 - => resolved on-the-fly to enrich with up-to-date background knowledge
 - e.g., every update in a lexical resource is available to every resource linked with it
 - => updates are available at query time
- Inconsistencies can be avoided through versioning

Ecosystem, infrastructure and community



- RDF and related standards are maintained by an active and relatively large community
 - Different fields of application
 - Libraries, GeoData, BioMed, ...
 - Established W3C standard and technological infrastructure
 - Linguistically relevant resources already provided
- RDF facilitates distributed development, re-using data, and, indirectly, interdisciplinary cooperation

Towards LLOD



- There are independent motivations to provide data in RDF
 - ❑ generic data model
 - ❑ generalization over heterogeneous DB schemes
 - ❑ generalization over heterogeneous terminologies
 - ❑ connect existing resource portals
 - ❑ a conceptual view on annotations

Towards LLOD



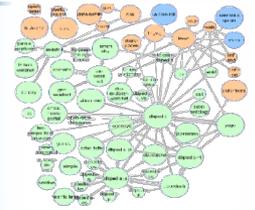
- There are independent motivations to provide data in RDF
- Side-effects: Data can be
 - linked with other RDF resources
 - e.g., lexicon – linguistic terminology
 - queried (federation)
 - even if stored in physically separated repositories

Towards LLOD



- There are independent motivations to provide data in RDF
 - Side-effects: Data can be
 - linked with other RDF resources
 - queried (federation)
- ⇒ Linguistic Linked Open Data (LLOD)
- An on-going effort orchestrated by the Open Linguistics Working Group of the Open Knowledge Foundation

Building the Cloud: Examples



- Each data provider has different incentives to use Linked Data and/or RDF
- Concepts of RDF and Linked Data have been brought up to solve open problems in different subcommunities of linguistics and neighboring fields

• Examples

- Corpora

Cassidy (2010), Chiarcos (2012)

- Machine-readable dictionaries

- Term and data bases

Chiarcos (2010)

- Etymological dictionaries

Chiarcos & Sukhareva (2014)

- combined queries

Burchardt et al. (2008), Rehm et al. (2008), Chiarcos & Götze (2007)

Linked Germanic Etymologies

Wiktionary [wik|farr] n., a wiki-based Open Content dictionary

swelt

Contents (hide)

- 1.1 Pronunciation
- 1.2 Etymology 1
- 1.3 Etymology 2
- 1.3.1 Verb
- 1.4 Anagrams

English [edit]

Pronunciation [edit]

- IPA (key): /swɛlt/

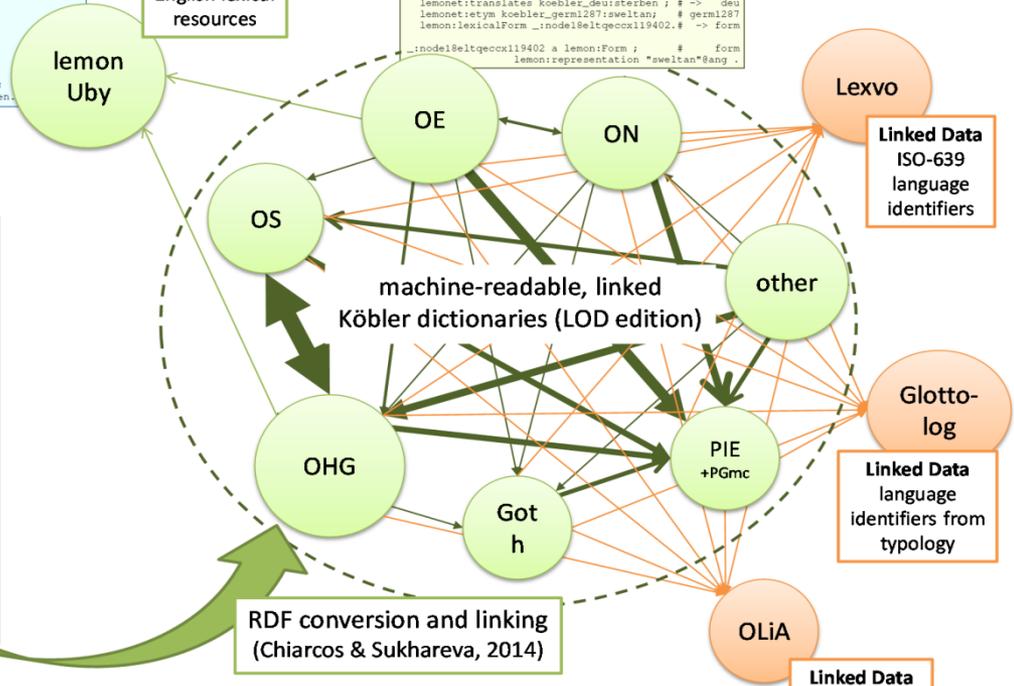
Etymology 1 [edit]

Old English *sweltan*. Cognate to Dutch *zwellen* ('to die').

```
# lexical entry
wen:WktEM_LexicalEntry_100154 a lemon:LexicalEntry
# sense
  lemon:sense wen:WktEM_sense_158899;
# form
  lemon:canonicalForm wenLE100154:CanonicalForm.
# form
  wenLE100154:CanonicalForm a lemon:Form;
  lemon:writtenRep "swelt"@en.
# sense
  wen:WktEM_sense_158899 a lemon:LexicalSense;
# sense definition
  lemon:definition :Statement1.
# sense definition for "swelt"
:Statement1 a lemon:SenseDefinition;
  uby:statementType "etymology";
  lemon:value "(obsolete) To die."@en;
  lemon:value "Old English sweltan."@en.
```

Linked Data German & English lexical resources

```
<http://purl.org/acoli/lex/koebler/angsweltan>
  a lemon:Word;
  lemon:language "ae."@deu;
  lexvo:language <http://lexvo.org/id/iso639-3/ang>;
  lemonet:hasPos "st. V. (3b)";
  oia:isVerb;
  lemonet:hasMorph "swel-t-an".
# cross-references to German (deu), (Proto)-Germanic
# (germ1287) and form
<http://purl.org/acoli/lex/koebler/angsweltan>
  lemonet:translates koebler_deu:sterben;
  lemonet:etym koebler_germ1287:sweltan;
  lemon:lexicalForm _:node1&lt;geocx119402.&# -> form
  _:node1&lt;geocx119402 a lemon:Form;
  lemon:representation "sweltan"@ang .
```



RDF conversion and linking (Chiarcos & Sukhareva, 2014)

XML edition and DB interface

IHGEL Corpus of Interlingual Historical Germanic Etymological Lexica

sweltan swel-t-an, ae., at. V. (3b); nhd. sterben, unkommen; Übersetzungsgeschichte lat. interire, mori occumbere

Verweise
s. ä-, efenge-, efen-ge-, ge-;

Etymologie
germ. *sweltan, at. v., sterben;
s. idg. *suel- (2), v., schwelen, brennen, Pokorny 1045;

Literatur
Hh 335, Hall/Meritt 330b, Lehnert 198b

XML conversion (Price 2012)

original OE Köbler dictionary (human-readable PDF) (<http://www.koeblergerhard.de/germanistischewoerterbuecher/altenglischewoerterbuch/AENG-5.pdf>)

Linked Germanic Etymologies

Application: Aligning early medieval gospel harmonies

Old Low German
(Old Saxon)

ALIGNMENT OF HELIAND AND TATIAN

Old High
German

HELIAND

[gód uuord angegin]: **Tha**n gi **[god]** uuillea[n], [quað he], uueros mid iuuuon **uuordun** uualdand **grôtean**, allaro **cuningo** craftigostan, than queðad gi, sô ic iu lêriu: [Fadar **usa**] friho barno, [thu **bist**] an them hôhon **himila** rikea, geuuihid **si** thîn **namo** [uuordo **gehuuילו**]. [Cuma] **thîn** [craftag] riki. Uuerða thîn uuilleo obar thesa uuerold **[alla]** **sô** sama an erðo, sô thar uppa ist an them hôhon [himilo rikea]. Gef ús dago gehuulikes rād, **drohtin** the gôdo, thîna hêlaga [helpa], endi alāt ús, hebenes uuard, managoro [mên]sculdio, al sô uue ôðrum mannum dôan. Ne lāt ús farlêdean **leða** uuihti sô forð an iro uuilleon, sô uui uuirðige sind, ac help ús uuiðar allun ubilon dâdiun. Sô sculun [gi] biddean, than gi te bede hnîgad uueros **mid iuuuon uuordun**, that iu uualdand god lêðes alâte an leutcunnea. Ef gi than uuilliad alâtan liudeo gehuuilicun thero sacono endi thero sundeono, the sie uuið iu selbon hîr **uurêða** geuuirkeat, **than** alâtîd **iu uualdand** god, **fadar** alamahtig **firnuuerk** mikil, managoro [mên]sculdeo. Ef iu than uuirðîd **iuuua môd te starc**, that gi ne uuileat ôðrun erlun **alâtan**, uueron uuamdâdi, than ne uuil iu ôc uualdand **god** grimuuerc **fargeban**, ac **gi** sculun **is** **geld** niman, suiðo lêðlic lôn te languru huuילו, alles thes unrehtes, thes gi ôðrum hîr gilêstead **an** thesumu **liohte** endi **than** **uuið** liudeo barn

TATIAN

The **quad** her in: **thanne** ir betot, thanne quedet sús: fater unser **thu thar bist in** himile, **si** giheilagot thîn **namo**, **queme** thîn rihhi, **si** thîn **uuiilo**, **sô** hêr in himile ist, sô si **hêr** in erdu, unsar brôt tagalihhaz **gib** uns **hiutu**, inti **furlaz** uns unsara sculd Oba ir furlazet mannum iro **sunta**, thanne furlazit iu iuuar fater thie **himilisco** iuuaara sunta. Oba ir ni furlazet mannum, thanne **ni** furlazit iu iuuar fater iuuaara sunta.

PAGE NUMBER: 63

Select measures:

- Relative Position? 0.17
- Relative Levenstein Distance? 0.4
- Identical Words?
- Dictionary?

Apply operations:

- Addition
- Multiplication? 0.63

Apply KNN

K: 90
N: 700
T: 0.07

Linked Germanic Etymologies

Conversion of etymological dictionaries to RDF

- **Linkability:** representation of relations within and beyond lexicons
- **Interoperability:** (meta)data representation through community-maintained vocabularies (lexvo, Glottolog, OLiA, *lemon*)
- **Inference:** filling the logical gaps of the original XML representation
 - *Symmetric closure of cross-references*

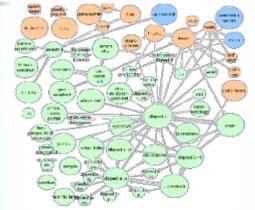
Linked Germanic Etymologies

lexicon	West Germanic				other			reconstr.	
	OE	OHG	OS	OLF	OFr	ON	Got	PGmc	PIE
entries (XML, in K)	25	24	9	2	13	12	5	9	7
triples (RDF, in M)	1.2	1.6	.6	.2	.6	.7	.4	.2	.2
lemon:Words & links (in K)									
OE	25						1		
OHG	2	26	7	2	3	1			
OS	1	4	9	1	2	1			
ON	1				1	14			
Got	1	1			1	1	6		
PGmc	5	3	3	1	2	4	2	8	
PIE	2	1	1	1	1	1	1		8
German	16	23	8	4	10	12	7	6	3
English		10	4	2	5		9		2
symmetric closure of etym. links (triples <i>per lang.</i> in K)									
	+11	+14	+11	+5	+9	+8	+5	+21	+9
links to (L)LOD data sets (triples <i>per data set</i> in K)									
OLiA	24	22	8	2	12	11	5	8	7
lexvo	132	186	82	21	68	82	49	14	15
Glottolog	15	11	8	3	7	11	6	9	13

A few critical remarks

- existing vocabularies insufficient
 - lemon: etymological relations between lexical entries ?
 - glottolog/lexvo/ISO-693: reconstructed languages ?
 - OLiA: Old Germanic inflection paradigms ?

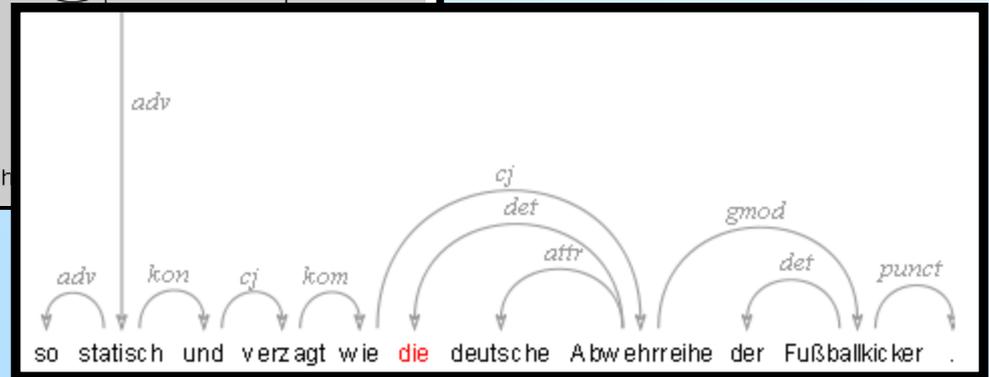
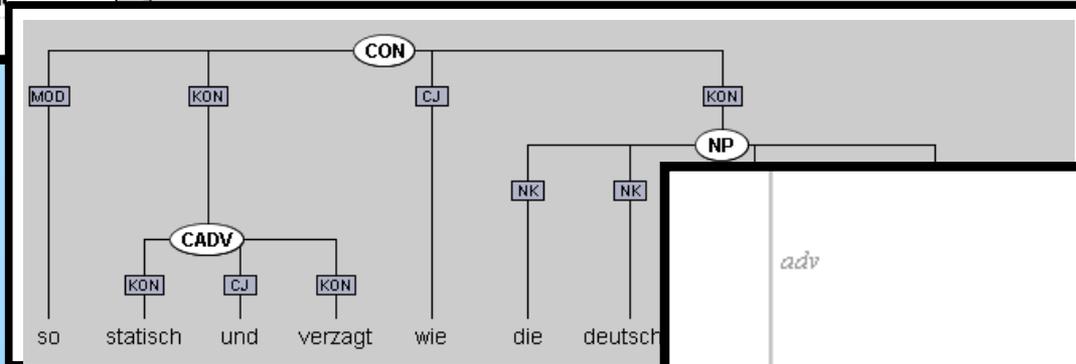
Use Case I: Corpora



Analyses produced by different researchers / NLP tools use different representation formalisms

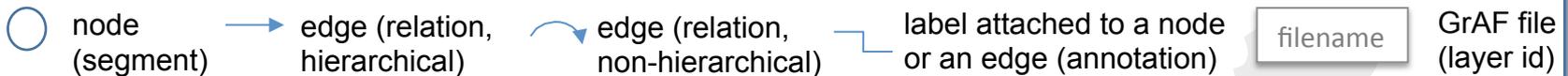
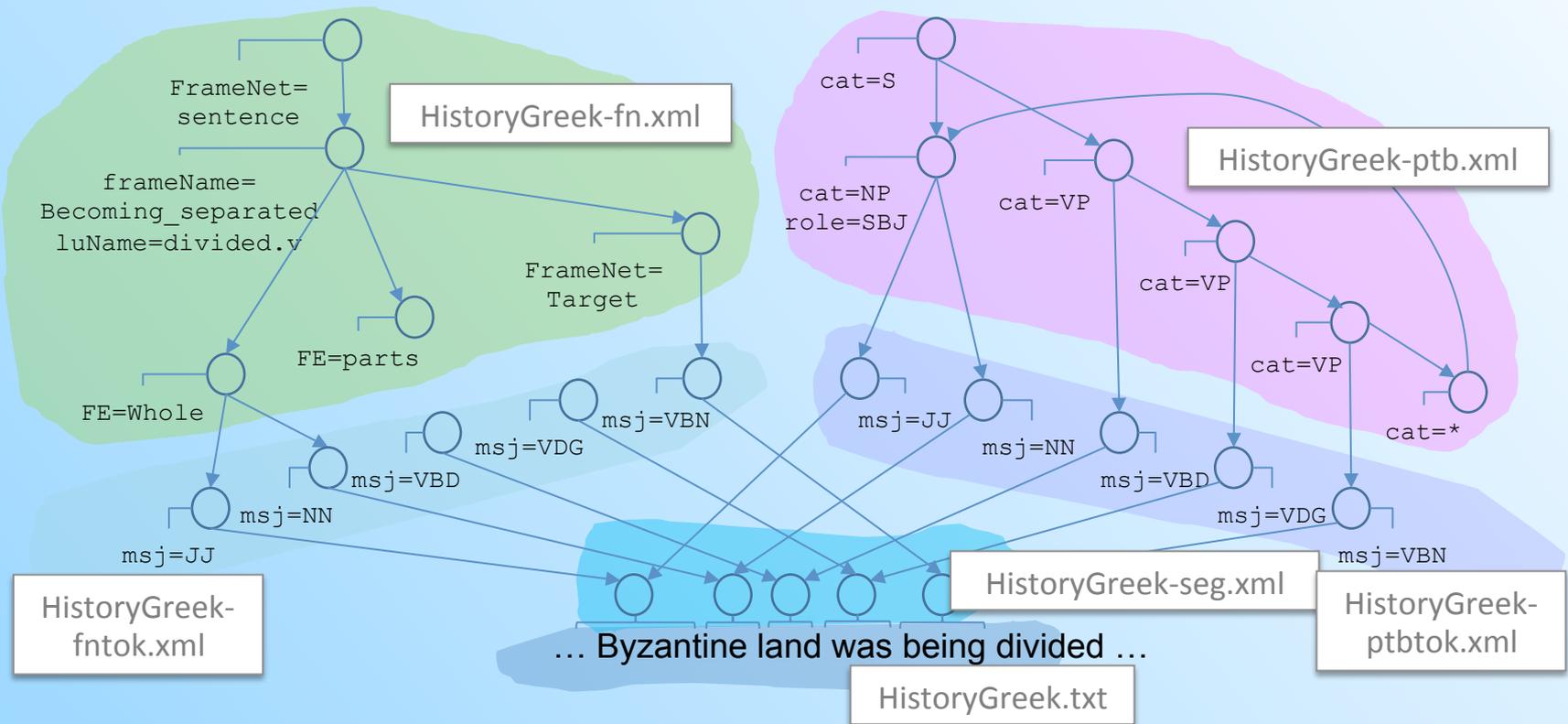
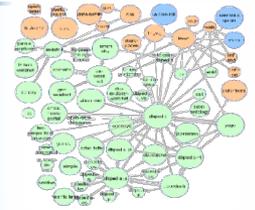
so statisch und verzagt wie **die** deutsche Abwehrreihe der Fußballkicker .
so statisch und verzagt wie der deutsch Abwehrreihe der Fußballkicker .
-- Pos -- Pos -- Nom.Sg.Fem Pos.Nom.Sg.Fem Nom.Sg.Fem Gen.Pl.Masc Gen.Pl.Masc --

Focus_newInf	nf-unsol						
Inf-Stat							acc-inf
NP							NP
Sent	s						
tok							Fußballkicker .

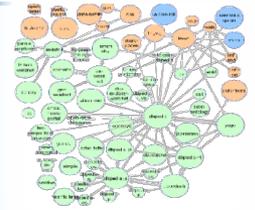


Standoff XML: GrAF, MASC corpus

(Ide & Suderman 2007, Ide et al. 2008)



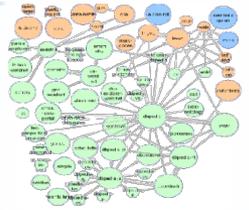
Use Case I: Corpora



Developmental trajectory

- proprietary/ad hoc formats
- XML
- directed acyclic graphs
 - standoff XML
- RDF
 - to store, query and integrate standoff annotations
 - Burchardt et al. (2008), Cassidy (2010), Chiarcos (2012), Rubiera et al. (2012), Hellmann et al. (2012), etc.

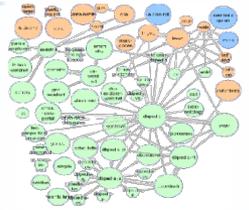
Use Case II: Lexical Resources



Similar trajectory

- proprietary formats
- XML, e.g., TEI dictionary
 - primarily for printed dictionaries
 - application in, e.g., NLP Budin et al. (2012)

Use Case II: Lexical Resources

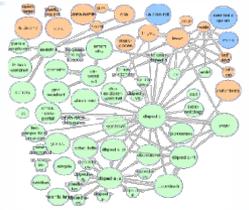


Similar trajectory

- proprietary formats
- XML, e.g., TEI dictionary
- generic data model, LMF
 - feature structures/directed acyclic graphs
 - linearized in XML/UML
 - **but:** semantics of links needs to be externally specified

(Francopoulo et al. 2009)

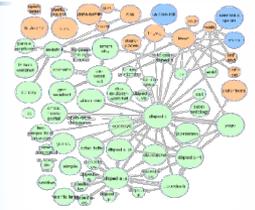
Use Case II: Lexical Resources



Similar trajectory

- proprietary formats
 - XML, e.g., TEI dictionary
 - generic data model, LMF
 - RDF, e.g., lemon
 - Buitelaar et al. (2013), Fiorelli et al. (2013), Moran & Brümmer (2013)
- => links to language resources other than lexicons

Use Case III: Bring them together



e.g.,

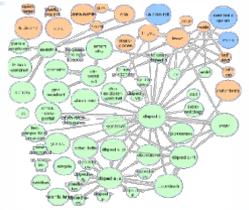
- words in a corpus and lexical concepts
 - at the moment, there exists no other unified formalism capable to express the following query

WordNet syntax WordNet

all hyponyms of *land* modified by a pertainym of *Byzantium*

- lemmas in a lexicon and grammatical features from a repository of linguistic reference terminology

Use Case IV: Term Bases



- linguistic terminology
 - Ontologies of Linguistic Annotation (<http://purl.org/olia>)
- language identifiers and descriptions
 - Glottolog (<http://glottolog.org>)

=>

- ontology-based querying
 - via query rewriting for corpus information systems
- ensemble combination architectures
 - merge output of different NLP tools on a conceptual basis

(Rehm et al. 2007, Chiarcos & Götze 2007)

(Chiarcos 2010, Pareja-Lora 2010)