SyPhy

A Phylosyntactic Approach to Model the Evolution of Indo European Languages

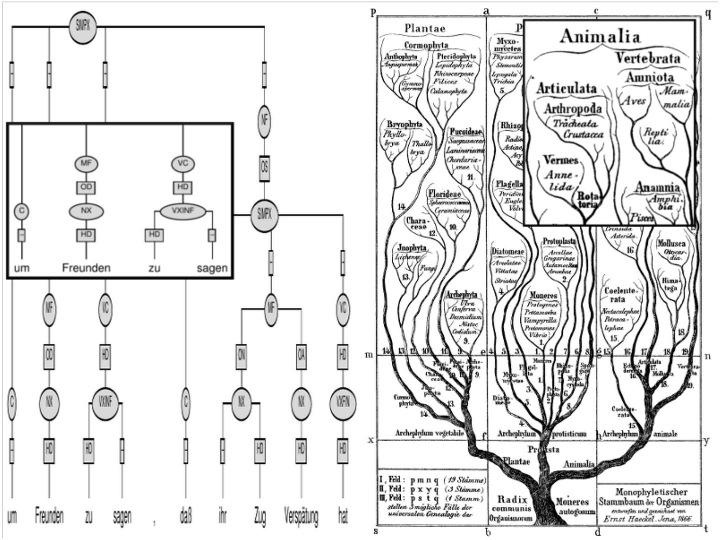
(Mid Term Presentation)

Markus Ackermann

Universität Leipzig

May 27 2014

A Tale of Trees



<span style="color:blue">left:</span> phrase structure tree from the Tüba-D/Z corpus (Telljohann et al. 2012)
right: one of the first monophyletic tree of life from Haeckel 1866

Outline

1. Materials and Research Questions
   - Corpora
   - Research Questions

2. Approaches and Methods
   - Data Preparation
   - Analyses

3. Outlook

4. Contribution to (Humanities) Research

Corpora

- several syntax-annotated corpora will be used
- part-of-speech classifications for token and analyses of syntax structure of the sentences
- focus on corpora adopting the constituency/phrase structure paradigm for syntax analyses for now

Name	**Language**	**Size**	**License**
YTH-PC-OEP[1]	Old English	1,5 mio. token	prop.; non-comercial
TüBa-D/Z	German	1,5 mio. token	prop.; academic research
CINTIL	Portugese	110,000 token	prop.; academic resarch
French Treebank	French	1 mio. token	prop; academic research
IcePaHC[2]	Icelandic	1 mio. words	LGPL

[1] The York-Toronto-Helsinki Parsed Corpus of Old English prose

[2] Icelandic Parsed Historical Corpus

Research Questions

1. Can a common ancestry of Indo-European languages be reconstructed looking at syntactic features of IE languages?
2. Which syntactic features that can be derived from POS and syntax annotations are most informative and distinctive for IE languages?
3. Will phylo-syntactic analyses yield genealogical trees for the IE languages in agreement with trees previously obtained using other methods?

Research Questions

1. Can a common ancestry of Indo-European languages be reconstructed looking at syntactic features of IE languages?
2. Which syntactic features that can be derived from POS and syntax annotations are most informative and distinctive for IE languages?
3. Will phylo-syntactic analyses yield genealogical trees for the IE languages in agreement with trees previously obtained using other methods?

personal motivation:

Investigations on Syntax are interesting.

Research Questions

1. Can a common ancestry of Indo-European languages be reconstructed looking at syntactic features of IE languages?
2. Which syntactic features that can be derived from POS and syntax annotations are most informative and distinctive for IE languages?
3. Will phylo-syntactic analyses yield genealogical trees for the IE languages in agreement with trees previously obtained using other methods?

personal motivation:

Investigations on Syntax are interesting.

Computational methods inspired by life sciences are interesting.

Research Questions

1. Can a common ancestry of Indo-European languages be reconstructed looking at syntactic features of IE languages?
2. Which syntactic features that can be derived from POS and syntax annotations are most informative and distinctive for IE languages?
3. Will phylo-syntactic analyses yield genealogical trees for the IE languages in agreement with trees previously obtained using other methods?

personal motivation:

Investigations on Syntax are interesting.

Computational methods inspired by life sciences are interesting.

A combination of both = really cool topic

Research Questions

- phylolinguistic analyses can help to further the understanding of early human history
  - ⇒ results can confirm or falsify hypotheses drawn from archeologic evidence, evolutionary anthropology and historical linguistics
  - ⇒ Gray and Atkinson 2003: phylolinguistic analysis of IE lanugages supports the Anatolian farmer hypothesis
  - ⇒ approach was based on lexical properties of the IE lanugages; additional phylosyntactic analyses for the IE family could in turn confirm or falsify Gray & Atkinsons investigations

Research Questions

- phylolinguistic analyses can help to further the understanding of early human history
  - ⇒ results can confirm or falsify hypotheses drawn from archeologic evidence, evolutionary anthropology and historical linguistics
  - ⇒ Gray and Atkinson 2003: phylolinguistic analysis of IE lanugages supports the Anatolian farmer hypothesis
  - ⇒ approach was based on lexical properties of the IE lanugages; additional phylosyntactic analyses for the IE family could in turn confirm or falsify Gray & Atkinsons investigations
- application of phylogenetic methods already help answering philological questions on text heritage
  - ⇒ Barbrook et al. 1998: phylogenetic analyses of 58 manuscripts of a Canterbury Tales story indicate distinct ancestry for groups of these manuscripts
  - ⇒ activities in SyPhy helps building up experience for possibly more syntactically informed phylolinguistic analyses

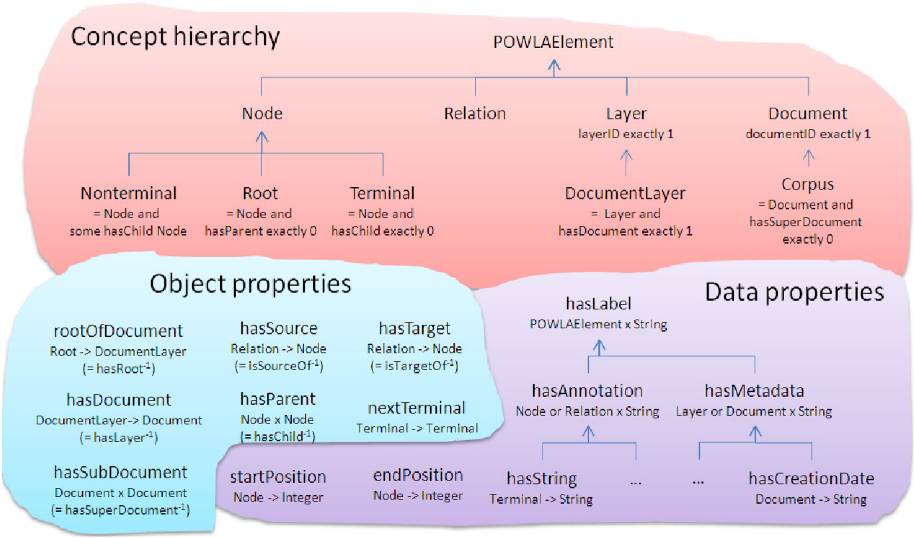
Data Preparation - Locating Data Sources

- search: query of the Clarin VLO (Uytvanck et al. 2010), the Oxford Text Archive and Google, hints by colleagues
- search criteria: Indo European language, syntax annotations (preferrably at least manually revised and using constituency/phrase grammar model)
- itemization of corpora with relevant properties:
size, annotion method, syntax model (dependency vs. constituents)
accessibility/licensing, format of corpus data

Data Preparation - Data Integration

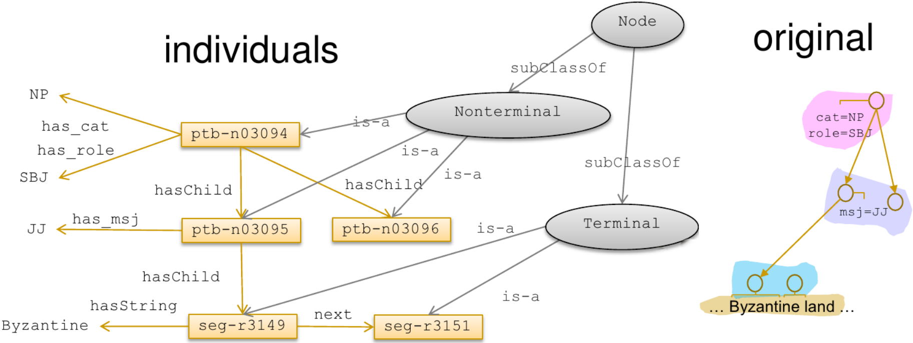
- adoption of apporach by Chiarcos 2012: POWLA
- representation of corpus data in RDF

Data Preparation - Data Integration



overview for POWLA Top ontology (Chiarcos 2012)

Data Preparation - Data Integration



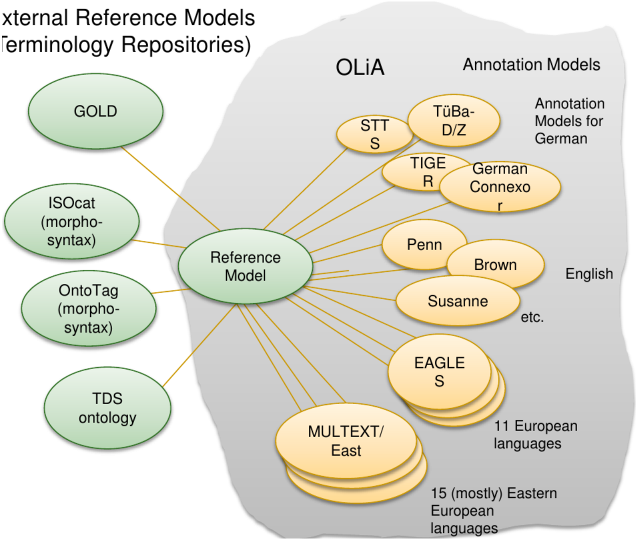
example for linguistic annotations in POWLA (Chiarcos 2012)

Data Preparation - Data Integration

- adoption of apporach by Chiarcos 2012: POWLA
- representation of corpus data in RDF

- use unifying domain ontology: OLiA (Ontology for Linguistic Annotations, Chiarcos 2008)
- unification/linking of different denotations for part of speech and constituent types via OLiA link ontologies (provided for some corpus formats, will be created by ourselves for others)

Data Preparation - Data Integration



import relations and usage of OLiA (Chiarcos 2008)

Data Preparation - Data Integration

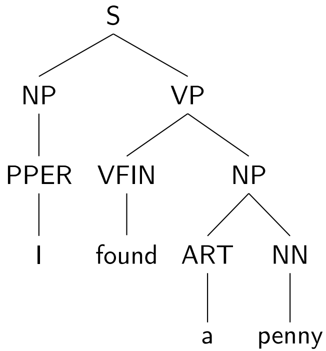
- adoption of apporach by Chiarcos 2012: POWLA
- representation of corpus data in RDF
- use unifying domain ontology: OLiA (Ontology for Linguistic Annotations, Chiarcos 2008)
- unification/linking of different denotations for part of speech and constituent types via OLiA link ontologies (provided for some corpus formats, will be created by ourselves for others)

Result: language data with annotations...
...in unified data format
...that are conceptually interoperable (for constituency and POS)
...accessible using a single set of technologies
(SPARQL queries; RDF APIs; Gremlin via SAIL, ...)

Analyses - Overview

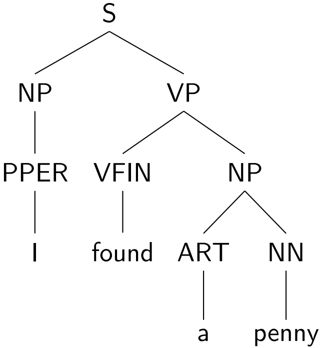
- extract syntactic features for each language from the corpora (frequencies syntax tree paths or patterns, pos ngrams, . . .)
- evaluate distinctivesness and applicability of extracted feature in a language classification setting ⇒ feature selection
- use resulting feature matrix for pyhlogenetic analysis
- re-iterations after error analyses or additions of further data sources

Analyses - Extracting Syntax Features



syntax tree paths:
S-NP-PPER
S-VP-VFIN
S-VP-NP-ART
S-VP-NP-NN

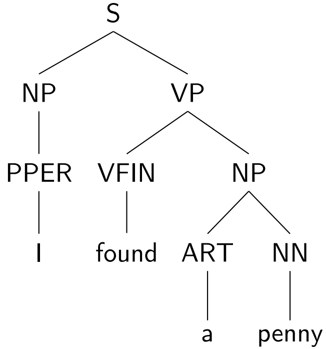
Analyses - Extracting Syntax Features



syntax tree paths:
S-NP-PPER
S-VP-VFIN
S-VP-NP-ART
S-VP-NP-NN

subtree patterns:
S > NP, VP
NP > PPER
NP > ART, NN
VP > VFIN, NP

Analyses - Extracting Syntax Features

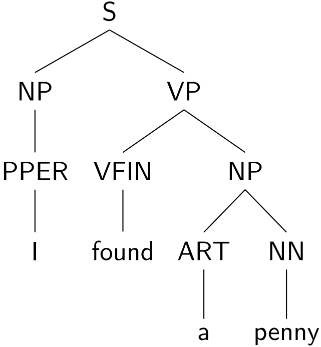


syntax tree paths:
S-NP-PPER
S-VP-VFIN
S-VP-NP-ART
S-VP-NP-NN

subtree patterns:
S > NP, VP
NP > PPER
NP > ART, NN
VP > VFIN, NP

part of speech bigrams:
$-PPER, PPER-VFIN, VFIN-ART, ART-NN, NN-$

Analyses - Extracting Syntax Features



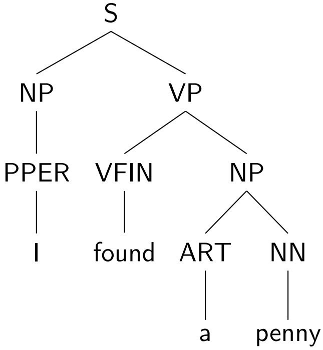
syntax tree paths:
S-NP-PPER
S-VP-VFIN
S-VP-NP-ART
S-VP-NP-NN

part of speech bigrams:
$-PPER, PPER-VFIN, VFIN-ART, ART-NN, NN-$

subtree patterns:
S > NP, VP
NP > PPER
NP > ART, NN
VP > VFIN, NP

dominated POS-seqs:
S :> PPER,VFIN,ART,NN
VP :> VFIN, ART, NN
NP :> PPER
NP :> ART, NN

Analyses - Extracting Syntax Features



syntax tree paths:
S-NP-PPER
S-VP-VFIN
S-VP-NP-ART
S-VP-NP-NN

part of speech bigrams:
$-PPER, PPER-VFIN, VFIN-ART, ART-NN, NN-$

subtree patterns:
S > NP, VP
NP > PPER
NP > ART, NN
VP > VFIN, NP

dominated POS-seqs:
S :> PPER,VFIN,ART,NN
VP :> VFIN, ART, NN
NP :> PPER
NP :> ART, NN

for each sentence: count occurence count for each feature and compute a feature score adapting the tf/idf idea

for each corpus: compute the mean feature vector over all sentences in the corpus to obtain a feature vector for the language as a whole

Analyses - Feature Evaluation and Feature Selection

- initially we will have a huge feature space, but limited example sentences to learn from

  ⇒ bias/variance problem likely

- whole feature classes or single feature manifestations will probably be uninformative, yielding more 'noise' than distinctive information

  hence: evaluate feature for their usefulness in the simpler setting of language classification before starting phylogenetic analyses

⇒ train log. regression classifiers with the syntax features applying Ridge regularisation

⇒ evaluate aptitude of possible feature subsets by measuring overlap of clustering results with actual language groups; use soft computing optimization (GA, PSO) to find (semi-)optimal feature subsets

Analyses - Phylogenetic analysis

phylogenetic analysis: generate an ancestral tree representing a hypothesis about the lineages of a group of species or genes based on agreement and differences in their characteristics

	(classic) phylogenetic analysis	**phylosyntactic analysis**
taxa	species, phenotypes, genomes	languages, texts, text groups
characteristics	physiological traits, abilities, behaviour patterns, amino acid sequences. . .	tree structures, POS sequences, dominance relations. . .
causes for mod.	changes in habitat, contact with new species, mutation, . . .	interaction with other cultures, discovery of new knowledge/technologies, new social structures, . . .

Analyses - Phylogenetic analysis

- several established methods to compute phylogenetic trees

distance matrix methods (DM): relies on explicitly provided distance measurements between the taxa and applies specialised clustering methods to devise a tree

+ direct control on distance estimation, fast (polynomial time)

– constant change rate assumption to produce trees with ancestry relationships

maximum parsimony (MP): search for potential phylogenetic trees explaining the data with a minimal number of evolutionary events (divergences, convergences, parallel evolutions)

+ in aggreement with Okham's razor (Parsimonieprinzip)

– computationally extensive (NP hard), heuristics required

maximum likelihood (ML): infers a probability distribution for possible tree according to the data (more evolutionary events in a tree make it less likely)

+ inherent MP-like effect, but more flexible (no constant rate assumptions needed)

– computationally expensive, appropriate change model required

Outlook

the next weeks: importers/converters from corpora to RDF/OWL; create and apply linking ontologies

at the end of the Vorlesungszeit: preliminary answers to questions (1) and (2) based on a language subset from 3-4 corpora

at the end of the semester: answer to question (3); refined anwers to (1) and (2) based on 2-3 additional languages

1. Can a common ancestry of Indo-European languages be reconstructed looking at syntactic features of IE languages?
2. Which syntactic features that can be derived from POS and syntax annotations are most informative and distinctive for IE languages?
3. Will phylo-syntactic analyses yield genealogical trees for the IE languages in agreement with trees previously obtained using other methods?

Outlook

the next weeks: importers/converters from corpora to RDF/OWL; create and apply linking ontologies

at the end of the Vorlesungszeit: preliminary answers to questions (1) and (2) based on a language subset from 3-4 corpora

at the end of the semester: answer to question (3); refined anwers to (1) and (2) based on 2-3 additional languages

possibilities for extension: integration of additional constituency based corpora; development of methods that allow also for dependency parsed corpora; try out additional syntax features; investigate influences of the phylognetic computation methods and their parameters

possibilies for reduction: intergration of fewer corpora; test less syntactic features;just use DM method for tree computation

Contribution to (Humanities) Research

methodological advancement

- additional insight in the advantages and limitations of characterising texts or languages by their syntactic characteristics for phylolinguistic analyses
- indication which kind of syntactic fingerprinting is most useful for studies on the ancestry of languages

Contribution to (Humanities) Research

methodological advancement

- additional insight in the advantages and limitations of characterising texts or languages by their syntactic characteristics for phylolinguistic analyses
- indication which kind of syntactic fingerprinting is most useful for studies on the ancestry of languages

contribution to Humanities

- additional support or falsification on an disputed aspect of the development of the IE language family
- methodoligal experiences and results will be (at least partially) transferable to pyhlosyntactic approaches for authorship attribution and stylometry (for syntax-analysed material) ⇒ ground work for *PhyloPhilology*?
- integrated RDF meta-corpus will be made publicly available to the extend permissible by the licenses ⇒ contribution to the Linked Linguistic Data Cloud

Bibliography I

[1] Adrian C. Barbrook et al. "The phylogeny of The Canterbury Tales". In: Nature 394 (Aug. 1998), pp. 839–846.

[2] Christian Chiarcos. "An ontology for linguistic annotations". In: LDV Forum 23.1 (2008), pp. 1–16.

[3] Christian Chiarcos. "POWLA: Modeling Linguistic Corpora in OWL/DL". In: The Semantic Web: Research and Applications. Ed. by Elena Simperl et al. Vol. 4. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2012, pp. 225–239. ISBN: 978-3-642-30283-1.

[4] Russell D. Gray and Quentin D. Atkinson. "Language-tree divergence times support the Anatolian theory of Indo-European origin". In: Nature 426 (Feb. 2003), pp. 435–439.

[5] Ernst Haeckel. Generelle Morphologie der Organismen. Berlin: Reimer, 1866.

Bibliography II

[6] Heike Telljohann et al.
Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z). Jan. 2012. URL: http://www.sfs.uni-tuebingen.de/fileadmin/static/ascl/resources/tuebadz-stylebook-1201.pdf (visited on 05/24/2014).

[7] Dieter Van Uytvanck et al. "Virtual Language Observatory: the portal to the language resources and technology universe". In: LREC 7 Proceedings. 2010, pp. 900–903.