HistoBankVis: Visualizing Language Change

Christin Schätzle and Miriam Butt

ICEHL XX

Workshop on Visualisations in Historical Linguistics
The University of Edinburgh

August 30, 2018

SFB-TRR 161
Quantitative Methods for Visual Computing
Transregional Collaborative Research Center

Universität
Konstanz

Research context

- SFB-TRR 161 “Quantitative Methods for Visual Computing”
- **Project D02 “Evaluation Metrics for Visual Analytics in Linguistics”** (Christin Schätzle)
  - Language change in Germanic and Indo-Aryan
  - How useful are visual analytic approaches to linguistic data?
  - Which visual variables and representations are most effective for which kind of problem/type of data?
- Project A03: Identification of subspaces/patterns in larger amounts of high-dimensional data (Michael Hund, Frederik Dennig)

$\Longrightarrow$ Historical linguistic data is **high-dimensional** and contains **subspaces** (e.g., interacting factors, relevant time periods) which need to be identified and understood.

DFG Deutsche Forschungsgemeinschaft

Visual Analytics for Linguistics (LingVis)

SFB-TRR 161

Les IV DECLINAISONS
La I Declinaison par le moyen de l'Article DEFINI
La II Declinaison par le moyen de l'Article INDEFINI
La III Declinaison par le moyen de l'Article d'UNITE
La IV Declinaison par le moyen de l'Article PARTITIF
Il n'a point de Plurier
Exemples.
Nouvelle GRAMMAIRE à l'usage des DAMES
CONJUGAISON de tous les Verbes réguliers
Participe actif
INFINITIF
Participe passif
1 Er, donner, aimer, chercher
2 Ir, batir, finir, punir, agir
3 Oir, devoir, recevoir, apercevoir
4 Re, Vendre, rendre, attendre
EXACTEUR ou IMPERATIF
Le Verbe affistant AVOIR
Le Verbe affistant ETRE
Temps Présent absolu.
Temps Présent dépendant
Temps passé stable
T. Futur conditionel
T. Futur absolu
T. passé historique absolu
T. passé histor. dépend.
Les Verbes Passifs se conjug. avec les Temps du Verbe Etre, tant simples que composés.

Paradigms visualized

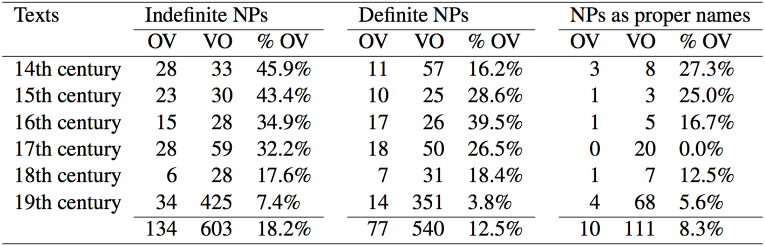
Acknowledgement and Thanks: **Frans Plank** originally inspired this LingVis enterprise!

Visual Analytics for Linguistics (LingVis)

- The Konstanz LingVis group to date has experimented with many different visualizations.
- Work by Christian Rohrdantz, Thomas Mayer, Dominik Sacha, Menna El-Assady, Annette Hautli-Janisz — see our websites
- But most of it
  - word-based
  - phonological and/or morphological features
  - simple intonation contours
- Currently trying to take things to a different level: syntax

Challenges for historical linguistics

'Traditional' approach: Pairwise comparison of the relevant information across a number of data tables with different characteristics



Texts	Indefinite NPs			Definite NPs			NPs as proper names		
	OV	VO	% OV	OV	VO	% OV	OV	VO	% OV
14th century	28	33	45.9%	11	57	16.2%	3	8	27.3%
15th century	23	30	43.4%	10	25	28.6%	1	3	25.0%
16th century	15	28	34.9%	17	26	39.5%	1	5	16.7%
17th century	28	59	32.2%	18	50	26.5%	0	20	0.0%
18th century	6	28	17.6%	7	31	18.4%	1	7	12.5%
19th century	34	425	7.4%	14	351	3.8%	4	68	5.6%
	134	603	18.2%	77	540	12.5%	10	111	8.3%

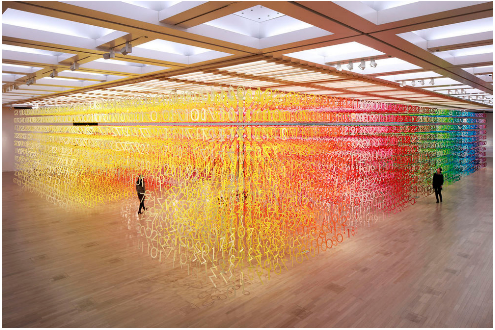
Definiteness distribution of NPs across different word orders in the history of Icelandic (Hróarsdóttir, 2000)

Challenges for historical linguistics

- Diachronic investigations involve understanding **highly complex interactions** between various linguistic and extra-linguistic features and structures, factoring in a **temporal dimension**.
- The factors underlying a change are often **unknown** or at least **highly debated** among researchers.
- **Data sparsity** may derogate statistical calculations.
- Interesting patterns may stay hidden when a researcher investigates temporal episodes that are either **too coarse or too fine grained**.

Meaningful patterns are difficult to see in the forest of numbers.

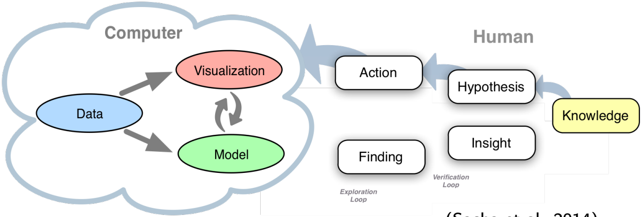
Visual Analytics for Linguistics (LingVis)



Emmanuelle Moureaux ‘Forest of Numbers’

General Aim: turn complex data sets and their relationships into at-a-glance visualizations complemented by the possibility to work interactively with different visual perspectives of the same complex relationships.

Visual Analytics for Linguistics (LingVis)



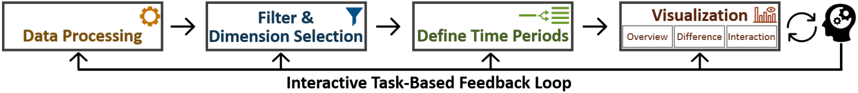
(Sacha et al. 2014)

Visual Analytics

- "Analyze first, show the important, zoom, filter and analyze further, details on demand" (Keim et al. 2008, based on Shneiderman 1996)
- Compact presentation of large amounts of data
- Different levels of detail on demand (interactivity)
- Exploratory and confirmatory data analysis
- Iterative process of hypothesis testing and generation

HistoBankVis: Visualizing language change

- Generically applicable system for historical linguistic research.
- Flexible investigation of a potentially high number of interacting linguistic features stored in an SQL database.



- Compact Matrix Visualization
  - Visualizes differences between selected dimensions across time
  - Measure of quality and “interestingness”
- Difference Histograms Visualization
- Dimension Interaction Visualization

Data Processing ⚙

- Concrete case study: interaction between subject case and word order in the history of Icelandic
- Reported word order changes in Icelandic:
  - change from OV to VO (Kiparsky 1996, Rögnvaldsson 1996, Hróarsdóttir 2000)
  - decrease of V1 (Franco 2008, Sigurðsson 1990, Butt et al. 2014)
- Research questions:
  - Which strategies are used to mark grammatical relations in Icelandic?
  - Do these strategies change diachronically?
  - Which functions do case and word order have at different stages of the language?

Icelandic Parsed Historical Corpus (IcePaHC)

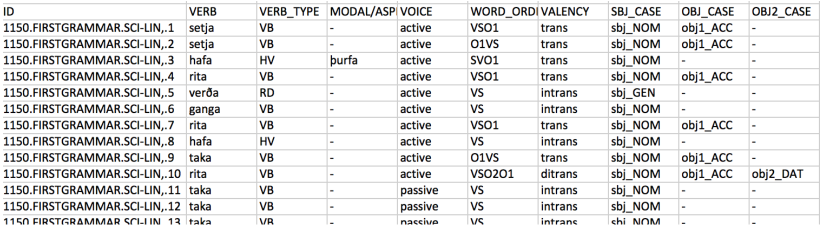
- 12th to 21st century – all attested stages of Icelandic.
- 61 texts, 1 million words, different genres (not representative across centuries).
- Annotation based on Penn Treebank style (Marcus et al. 1993).
- Information about sentence types, constituents, word order, grammatical relations, tense, voice, and case.

Sample IcePaHC Annotation

```
(IP-MAT-SPE (NP-SBJ (PRO-D Mér-mér))
  (VBPI finnst-finna)
  (CP-ADV-SPE (WADVP-1 0)
     (C sem-sem)
     (IP-SUB-SPE (ADVP *T*-1)
              (NP-SBJ (PRO-N ég-ég))
              (BEPS sé-vera) (VBN sloppinn-sleppa)
           (PP (P úr-úr) (NP (NP-POS (ONE+Q-G einhvers-einhver)
              (N-G konar-konar)) (N-D fangelsi-fangelsi)))))
     (.  .-.))
  (ID 1882.TORFHILDUR.NAR-FIC,.603))
```

Data Processing

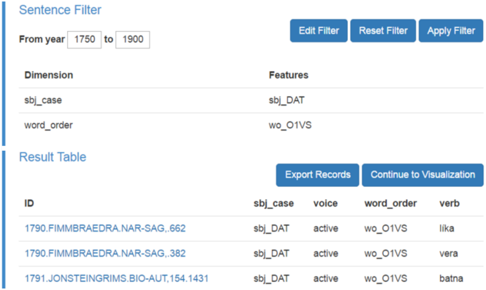
- Extraction of relevant linguistic data dimensions from the annotation of IcePaHC via Perl scripts
  → verb type, voice, word order, case and valency
- Information is collected for each matrix declarative sentence and mapped onto its sentence ID (gives information about the age, name, and genre of a text)
- Creation of well-structured CSV-file → data is stored in a relational SQL database in HistoBankVis



ID	VERB	VERB_TYPE	MODAL/ASPI	VOICE	WORD_ORDI	VALENCY	SBJ_CASE	OBJ_CASE	OBJ2_CASE
1150.FIRSTGRAMMAR.SCI-LIN,.1	setja	VB	-	active	VSO1	trans	sbj_NOM	obj1_ACC	-
1150.FIRSTGRAMMAR.SCI-LIN,.2	setja	VB	-	active	O1VS	trans	sbj_NOM	obj1_ACC	-
1150.FIRSTGRAMMAR.SCI-LIN,.3	hafa	HV	þurfa	active	SVO1	trans	sbj_NOM	-	-
1150.FIRSTGRAMMAR.SCI-LIN,.4	rita	VB	-	active	VSO1	trans	sbj_NOM	obj1_ACC	-
1150.FIRSTGRAMMAR.SCI-LIN,.5	verða	RD	-	active	VS	intrans	sbj_GEN	-	-
1150.FIRSTGRAMMAR.SCI-LIN,.6	ganga	VB	-	active	VS	intrans	sbj_NOM	-	-
1150.FIRSTGRAMMAR.SCI-LIN,.7	rita	VB	-	active	VSO1	trans	sbj_NOM	obj1_ACC	-
1150.FIRSTGRAMMAR.SCI-LIN,.8	hafa	HV	-	active	VS	intrans	sbj_NOM	-	-
1150.FIRSTGRAMMAR.SCI-LIN,.9	taka	VB	-	active	O1VS	trans	sbj_NOM	obj1_ACC	-
1150.FIRSTGRAMMAR.SCI-LIN,.10	rita	VB	-	active	VSO2O1	ditrans	sbj_NOM	obj1_ACC	obj2_DAT
1150.FIRSTGRAMMAR.SCI-LIN,.11	taka	VB	-	passive	VS	intrans	sbj_NOM	-	-
1150.FIRSTGRAMMAR.SCI-LIN,.12	taka	VB	-	passive	VS	intrans	sbj_NOM	-	-
1150.FIRSTGRAMMAR.SCI-LIN,.13	taka	VB	-	passive	VS	intrans	sbj_NOM	-	-

Task-based Filtering

- Explore dataset before visualization
- Construction of a task-specific dataset
  - Filter for sentences with relevant *features* (i.e., cell entries)
  - *Dimension* selection (i.e., columns)



Sentence Filter

From year 1750 to 1900

Dimension	Features
sbj_case	sbj_DAT
word_order	wo_O1VS

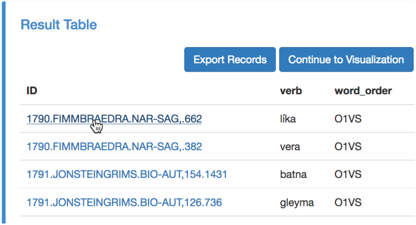
Result Table

ID	sbj_case	voice	word_order	verb
1790.FIMMBRAEDRA.NAR-SAG,.662	sbj_DAT	active	wo_O1VS	líka
1790.FIMMBRAEDRA.NAR-SAG,.382	sbj_DAT	active	wo_O1VS	vera
1791.JONSTEINGRIMS.BIO-AUT,154.1431	sbj_DAT	active	wo_O1VS	batna

⟹ Selected dimensions and features are analyzed in the visualization.

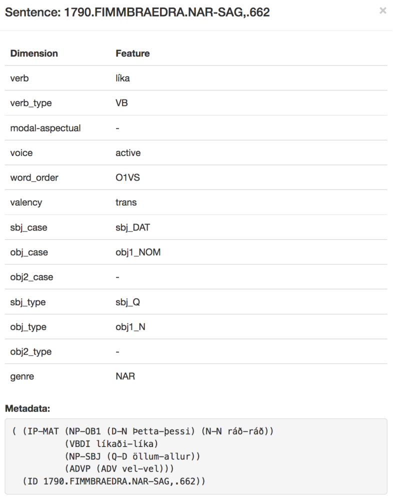
Task-based Filtering

- Access to detailed information about each data point
- Furthers understanding of data quality
- Comparison of annotated values and extracted features



Result Table

ID	verb	word_order
1790.FIMMBRAEDRA.NAR-SAG,.662	líka	O1VS
1790.FIMMBRAEDRA.NAR-SAG,.382	vera	O1VS
1791.JONSTEINGRIMS.BIO-AUT,154.1431	batna	O1VS
1791.JONSTEINGRIMS.BIO-AUT,126.736	gleyma	O1VS



Sentence: 1790.FIMMBRAEDRA.NAR-SAG,.662

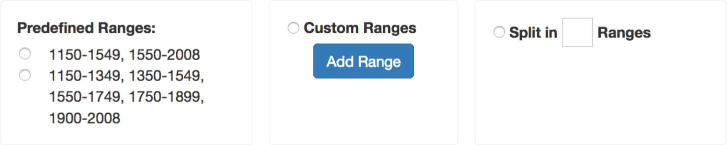
Dimension	Feature
verb	líka
verb_type	VB
modal-aspectual	-
voice	active
word_order	O1VS
valency	trans
sbj_case	sbj_DAT
obj_case	obj1_NOM
obj2_case	-
sbj_type	sbj_Q
obj_type	obj1_N
obj2_type	-
genre	NAR

Metadata:

```
( (IP-MAT (NP-OB1 (D-N Þetta-þessi) (N-N ráð-ráð))
          (VBDI líkaði-líka)
          (NP-SBJ (Q-D öllum-allur))
          (ADVP (ADV vel-vel)))
  (ID 1790.FIMMBRAEDRA.NAR-SAG,.662))
```

Analyzing Change over Time

- Define/select time periods



Predefined Ranges:

○ 1150-1549, 1550-2008

○ 1150-1349, 1350-1549, 1550-1749, 1750-1899, 1900-2008

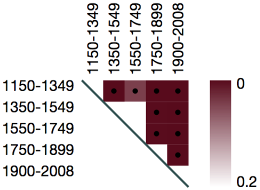
○ **Custom Ranges**

○ **Split in** □ **Ranges**

- Compact Matrix Visualization
- Difference Histograms Visualization
- Dimension Interaction Visualization

Compact Matrix Visualization

- Visualizes differences between selected dimensions across time
- Comparison of periods along the diagonal
- Differences mapped onto a colormap

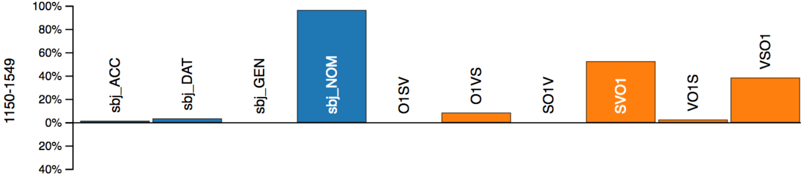


- Two comparison modes:
  - χ^2-test
    - Statistical significance ($\alpha \leq 0.05$) ●
    - Absence of necessary preconditions ✕
    - p-value is mapped to colormap (red $p = 0$, white $p \geq 0.2$)
  - Euclidean distance
    - Colormap indicates high (red) or low (white) distance
    - High Euclidean distance → large difference (high significance)

- Measure of quality and “interestingness”

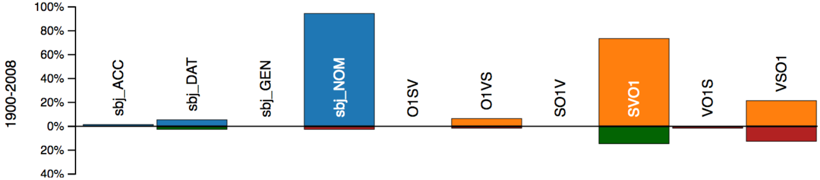
Difference Histograms Visualization

- Histograms provide detailed views on individual features and their diachrony.
- Each time period is visualized as one bar chart/histogram.
- Dimensions are encoded via different colors.
- Each bar in the histogram corresponds to an individual feature.
- The height of a bar shows the percentage of sentences containing the respective feature in the given time period.

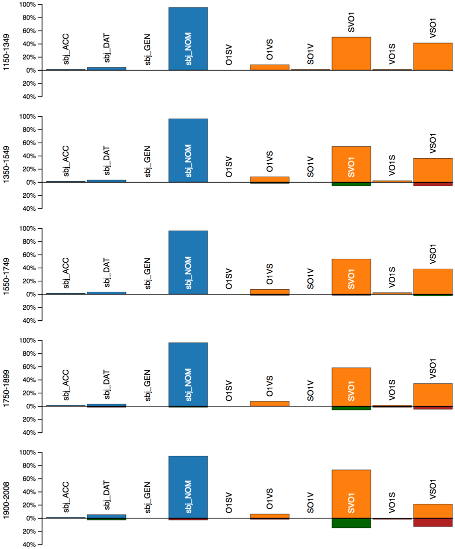


Difference Histograms Visualization

- Differences between periods are visualized as a separate bar chart below each bar:
    - green → feature increased
    - red → feature decreased
- Different comparison modes:
    - Previous period
    - First range
    - Last range
    - Average of all ranges
    - Average of previous ranges



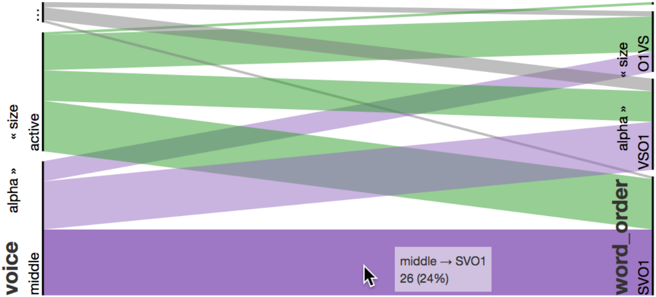
Difference Histograms Visualization



Dimension Interaction Visualization

SFB-TRR 161

- Application of the **Parallel Sets** technique (Bendix et al. 2005, Kosara et al 2006)
  - Each feature is visualized as a proportion of an equally spaced vertical line.
  - The vertical lines represent the data dimensions.
- Each time period is visualized as one Parallel Sets visualization.



Dimension Interaction Visualization

SFB-TRR 161

Parallel Sets

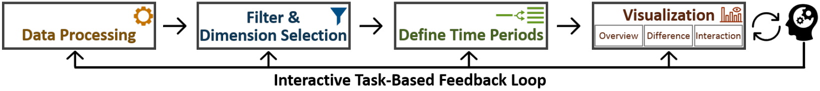
- allow for the flexible investigation of interactions between features from different data dimensions

  ⟶ **Dimension Interaction Visualization**

- Dimensions can be reordered via drag & drop.
- Features can be sorted according to size or alphabetically in an ascending or descending order.
- Mousing over a feature interaction provides information about the feature correspondence and the respective occurrence frequencies.

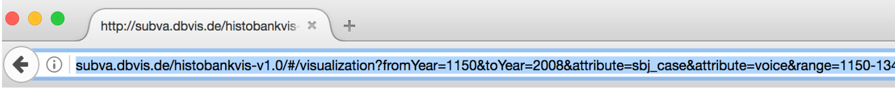
Hypothesis Generation and Feedback

- Generation and testing of new hypotheses
- Feed the knowledge gained back into the system:
  - Change feature filters
  - Select different dimensions
  - Use different time periods
  - Process data anew
- Iterative analysis process
- Combination of knowledge-based and data-driven modeling



Access and Usability

- On-line browser app: `http://histobankvis.dbvis.de/`
- Analysis steps and current views are encoded by unique identification URLs



http://subva.dbvis.de/histobankvis-

subva.dbvis.de/histobankvis-v1.0/#/visualization?fromYear=1150&toYear=2008&attribute=sbj_case&attribute=voice&range=1150-134

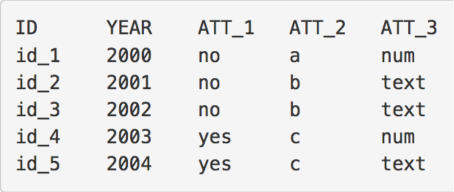
→ Store and retrieve visualizations/analyses

→ Share data and knowledge with other researchers

→ Supports research collaborations

Data upload

- IcePaHC dataset implemented as default
- Upload of own data
    - Tab-separated files
    - Must start with unique ID followed by a year date
    - Meta information, e.g., the corresponding full texts or parse trees, can be uploaded as well → unique IDs map between the files



```
ID      YEAR    ATT_1   ATT_2   ATT_3
id_1    2000    no      a       num
id_2    2001    no      b       text
id_3    2002    no      b       text
id_4    2003    yes     c       num
id_5    2004    yes     c       text
```

⟹ Further instructions are provided on-line!

Work in progress

- Investigation of the interrelation between case and word order in other Penn parsed corpora
- HeliPaD: a parsed corpus of Old Saxon (Walkden 2015)
- Penn Parsed Corpora of Historical English
  - York-Toronto-Helsinki Parsed Corpus of Old English prose (YCOE, Taylor et al. 2003)
  - Penn-Helsinki Parsed Corpus of Middle English, second edition (PPCME2, Kroch & Taylor 2000)
  - Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME, Kroch et al. 2004)
  - Penn-Helsinki Parsed Corpus of Modern British English (Kroch et al. 2010)

⟹ Test and improve data upload

⟹ Broaden scope of application of HistoBankVis

HistoBankVis – HeliPaD

Demo

`http://histobankvis.dbvis.de/`

(*Dimension Interactions not yet available on-line.)

Preliminary results – HeliPaD

- Dative objects are mostly pronouns, i.e., sentient/animate entities.
- Large tendency for animate dative arguments to precede the nominative argument.
- Yet, no diachronic perspective.

⟹ Penn Parsed Corpora of Historical English

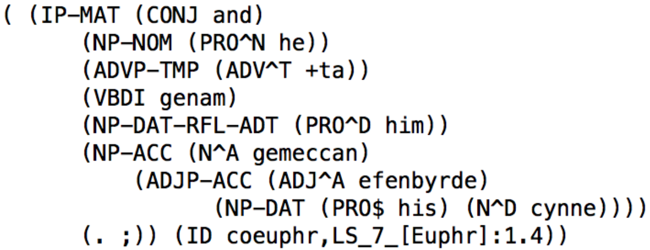
Penn Parsed Corpora of Historical English

- Problem: Corpora differ with respect to the annotation of grammatical relations and case marking (amongst other things)

- Lack of uniform standard (for Penn Treebanks overall)
- Difficult to automatically process the data

Issues of **reproducibility** and **comparability** of results!

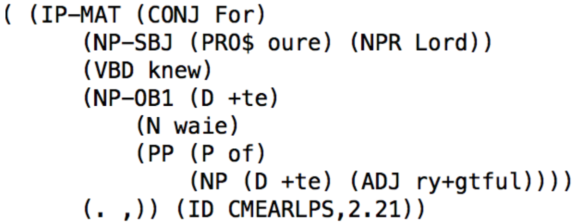
YCOE annotation (Old English)



```
( (IP-MAT (CONJ and)
      (NP-NOM (PRO^N he))
      (ADVP-TMP (ADV^T +ta))
      (VBDI genam)
      (NP-DAT-RFL-ADT (PRO^D him))
      (NP-ACC (N^A gemeccan)
          (ADJP-ACC (ADJ^A efenbyrde)
                (NP-DAT (PRO$ his) (N^D cynne))))
      (. ;)) (ID coeuphr,LS_7_[Euphr]:1.4))
```

Case marking, but no grammatical relations.

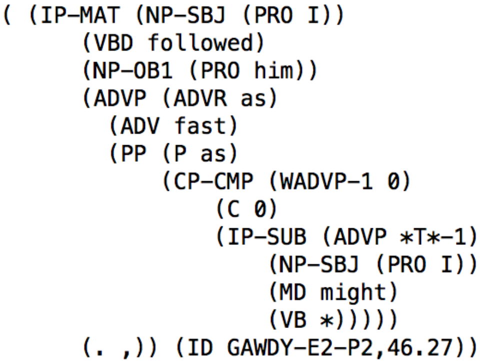
PPCME annotation (Middle English)



```
( (IP-MAT (CONJ For)
      (NP-SBJ (PRO$ oure) (NPR Lord))
      (VBD knew)
      (NP-OB1 (D +te)
          (N waie)
          (PP (P of)
              (NP (D +te) (ADJ ry+gtful))))
      (. ,)) (ID CMEARLPS,2.21))
```

No case marking, but grammatical relations.

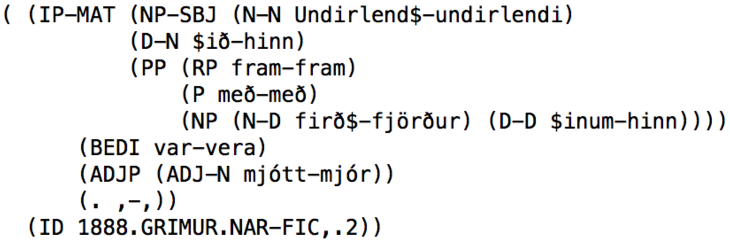
PPCEME2 annotation (Early Modern English)



```
( (IP-MAT (NP-SBJ (PRO I))
      (VBD followed)
      (NP-OB1 (PRO him))
      (ADVP (ADVR as)
        (ADV fast)
        (PP (P as)
            (CP-CMP (WADVP-1 0)
                (C 0)
                (IP-SUB (ADVP *T*-1)
                    (NP-SBJ (PRO I))
                    (MD might)
                    (VB *))))))
      (. ,)) (ID GAWDY-E2-P2,46.27))
```

No case marking, but grammatical relations.

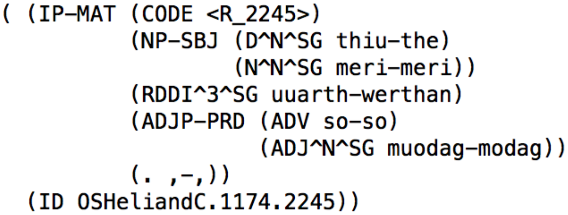
IcePaHC annotation



```
( (IP-MAT (NP-SBJ (N-N Undirlend$-undirlendi)
                  (D-N $ið-hinn)
                  (PP (RP fram-fram)
                      (P með-með)
                      (NP (N-D firð$-fjörður) (D-D $inum-hinn))))
              (BEDI var-vera)
              (ADJP (ADJ-N mjótt-mjór))
              (. ,-,))
  (ID 1888.GRIMUR.NAR-FIC,.2))
```

Case marking **and** grammatical relations (and lemmas). Yet, case is annotated differently than in the YCOE.

HeliPaD annotation



```
( (IP-MAT (CODE <R_2245>)
          (NP-SBJ (D^N^SG thiu-the)
                  (N^N^SG meri-meri))
          (RDDI^3^SG uuarth-werthan)
          (ADJP-PRD (ADV so-so)
                    (ADJ^N^SG muodag-modag))
          (. ,-,))
  (ID OSHeliandC.1174.2245))
```

Case marking **and** grammatical relations (and lemmas). Yet, case is again annotated differently.

Future Work

- Improve remaining flaws of data upload
- Automated solution for data processing
  - Integrate data processing into HistoBankVis pipeline
  - Build datasets via the filtering component directly from original corpus
  - Penn Treebanks and Universal Dendency Treebanks (CoNLL-format) as input
  - Develop standardized processing scheme
  - Integrate methods from the fields of data uncertainty and provenance
- Visual modeling of language change
  - Automatic identification of changing time periods
  - Automatically identify patterns of change, i.e., find the linguistic features involved in a change
  - S-curve model vs. cyclic patterns of change

Thank you!

Feedback? Suggestions for improvement?

`http://histobankvis.dbvis.de/`

christin.schaetzle@uni-konstanz.de
miriam.butt@uni-konstanz.de