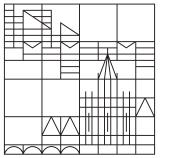
Universität Konstanz



Visual Analytics for Linguistics

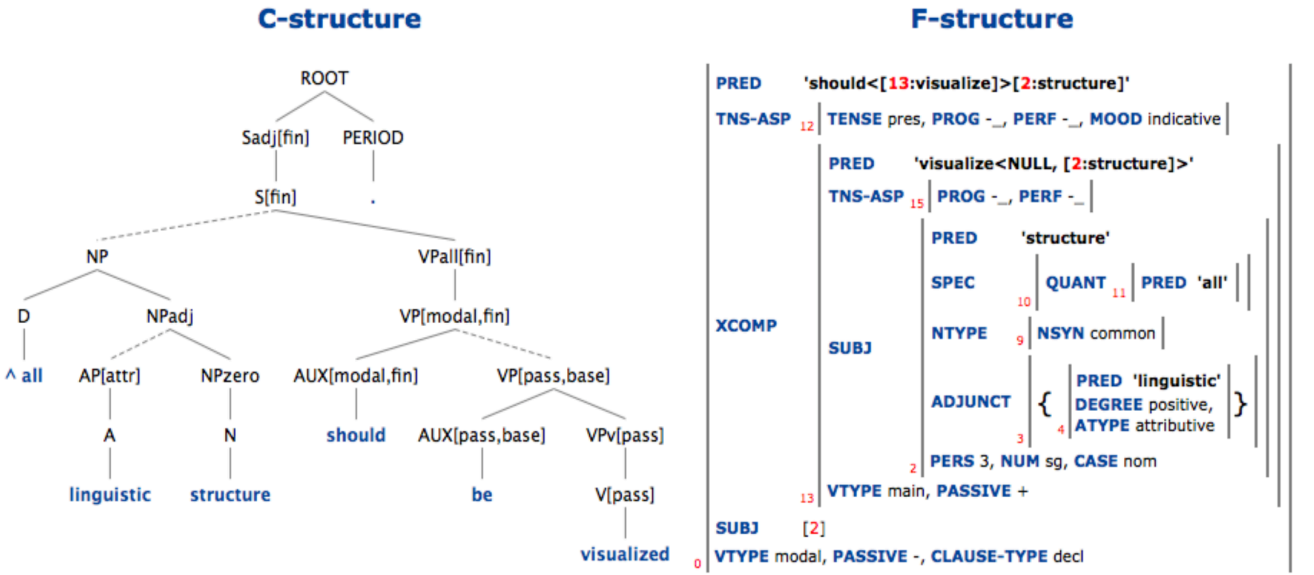
Christin Schätzle

University of Konstanz

Friday Lunch Talk
Linguistics Department
Yale University

Nov. 20, 2015

Standard Visualization: Syntax



Syntactic Analysis with Lexical-Functional Grammar (LFG)
`http://clarino.uib.no/iness/xle-web` (Web Interface for LFG Grammars)
Grammar developed at PARC

Mining Linguistic Data

Methodological Challenge/Opportunity

- Use of new technology to detect distributional patterns in language data.
- Ever increasing sources of digital data
  - Wikipedia, social media
  - Constructed corpora
- Specialized query and search tools (KWIC, COSMAS, DWDS, ANNIS)
- Programming languages specialized for text processing and statistical analysis (Python, Perl, R)

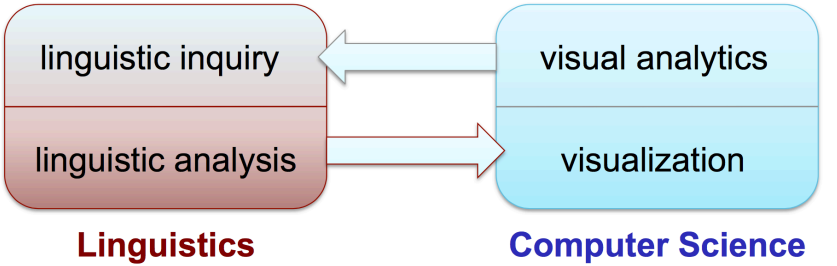
Problem: meaningful patterns are difficult to see in the forest of numbers

Opportunity: Visual Analytics for Linguistics (LingVis)

Visual Analytics for Linguistics (LingVis)

Overall interdisciplinary goal:

- Integrate methods from visual analytics into domains of linguistic inquiry
- Explore challenges based on the needs of linguistic analysis for visualization methods



- Iterations of hypothesis-formation and hypothesis-testing
- Exploratory visual access to data

Interdisciplinary Research Projects in Konstanz

VisArgue

Why and when do arguments win? Deliberation in political negotiations

Visual Analysis of Language Change and Use Patterns

Language change, genetic relationships between languages and variations in language use across time

Evaluation Metrics for Visual Analytics in Linguistics

Usefulness of VA within Linguistics

VisArgue - Analyzing Political Argumentation

Motivation:

- Large and expensive (public) projects lead to conflicts in society and politics.
- High risk for decision makers

  $\longrightarrow$ More knowledge needed on how consensus is achieved in discourse

Concept of deliberation (Habermas 1981, 1991):

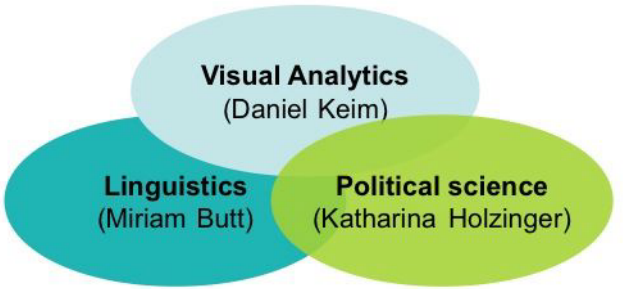
Stakeholders in a multilog...

- ... should justify their positions truthfully and rationally.
- ... should eventually defer to the better argument.

VisArgue - Analyzing Political Argumentation

SPONSORED BY THE
Federal Ministry of Education and Research

- **Aim:** Linguistically-motivated visual analysis of deliberation in political communication

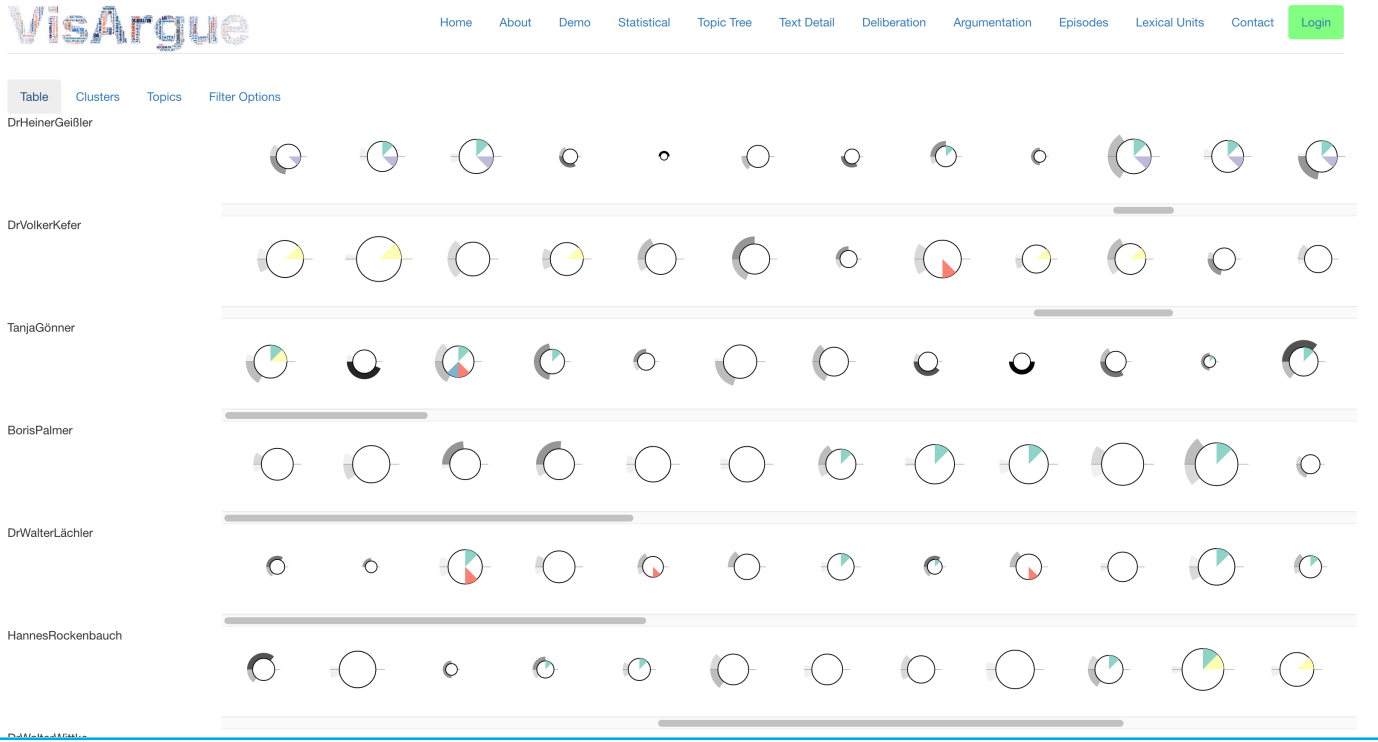


- Interdisciplinary project (University of Konstanz) as part of the BMBF *eHumanities* initiative in Germany

Data

- Transcribed minutes of the public arbitration for “Stuttgart 21” (Railway and urban development project in the city of Stuttgart)
  - 9 days of session, around 65 hours of discussion, 70 speakers
  - 1330 utterances, 265.000 tokens
  - rule-based annotation of utterances (hand-crafted rules)
- Speakers are either Pro or Contra
- Mediator is supposed to be neutral

Glyph Visualization of Utterance Context (El-Assady et al. submitted)



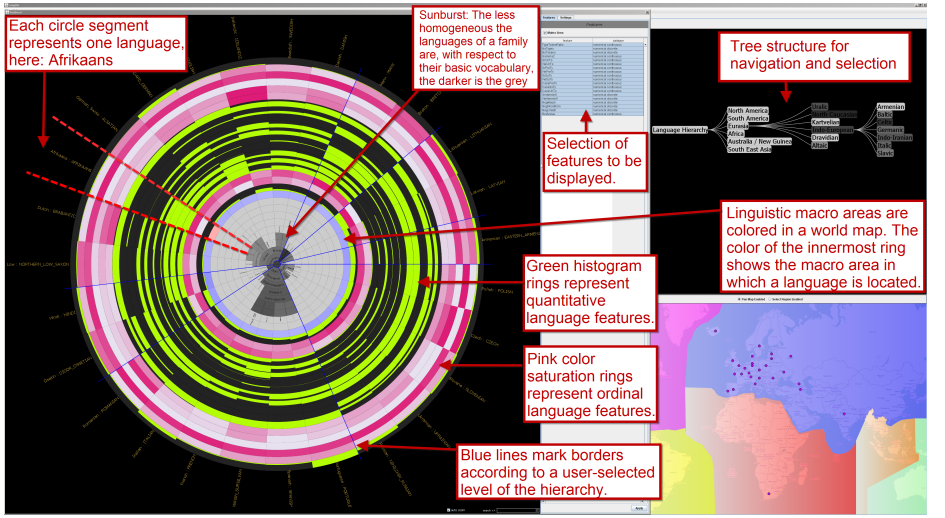
Visual Analysis of Language Change and Use Patterns

- **Aim:** Analysis of patterns of language change and use via the combination of linguistic analysis and novel visualization techniques
- Uncover information about...
  - language change
  - genetic relationships between languages
  - variations in language use across time
- Interdisciplinary cooperation (University of Konstanz) funded by the DFG (Deutsche Forschungsgesellschaft)
  - Linguistics (Miriam Butt)
  - Visual Analytics (Daniel Keim)

DFG Deutsche Forschungsgemeinschaft

The World's Language Explorer (Rohrdantz et al. 2012, Mayer et al. 2014)

- Sunburst visualization in conjunction with geo-spatial information (maps)
- Comparison of automatically extracted language features
- Interactive investigation of (dis)similarity of features across languages
- Online: WALS Sunburst Explorer `http://th-mayer.de/wals/#30A/`



Identifying N-V complex predicates in Hindi/Urdu (Butt et al. 2012)

```
#this file lists X in X+kar, X+ho, X+hu, X+rakh sequences with corr...
esponding occurrences in the (candidate) CP sequences
#X = word occurring directly to the left of LV (LV: kar, ho, hu, rakh)
#kar: # of occurrences of X with kar
#ho: # of occurrences of X with ho
#hu: # of occurrences of X with hu
#rakh: # of occurrences of X with rakh
X          #hu  #kar #ho  #rakh
اضافہ      674  466  524      0
شروع       378  2336 1691     0
معلوم      366  254  609      0
دھماکہ     359  135  44       0
حملہ       227  1232 100      0
متاثر      183  178  765      0
نقصان      173  0    114      0
ایک        172  373  7027     0
تبادلہ     147  394  588      0
تقریر      142  105  235      9
پیدا       103  754  956      0
کالعدم     102  1501 3609     0
برآمد      80   210  96       0
اعلان      74   0    263      0
تعمیر      62   59   1161     0
آغاز       59   315  75       0
ہی         56   0    2267     0
مذاکرات    54   197  262      0
انکشاف     51   165  13       0
```

- Identification of sequences of noun+verb for understanding complex predicate patterns in Hindi/Urdu, e.g. *phone-do, use-do, memory-come, begin-do/come*
- Data: 7.0 million word raw (unannotated) corpus of Urdu (BBC Urdu)

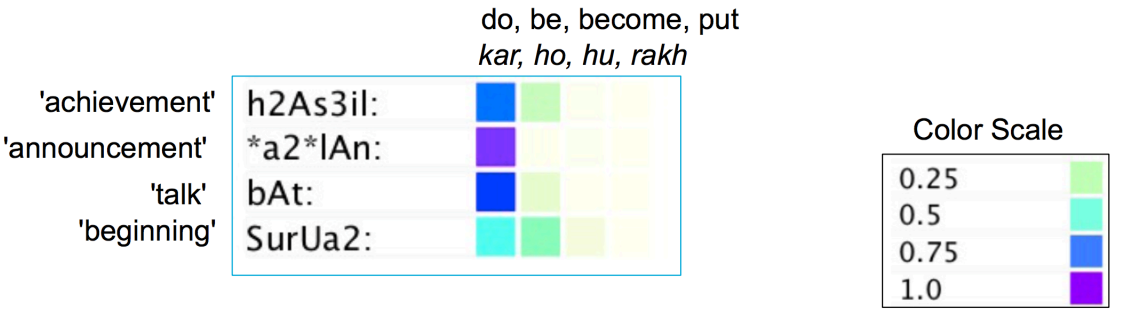
Identifying N-V complex predicates in Hindi/Urdu

Pixel visualization

Statistical Data:

ID	Noun	Rel. freq. with *kar*	Rel. freq. with *ho*	Rel. freq. with *hu*	Rel. freq. with *rakh*
1	حاصل	0.771	0.222	0.007	0.000
2	اعلان	0.982	0.011	0.007	0.000
3	بات	0.853	0.147	0.000	0.000
4	شروع	0.530	0.384	0.086	0.000

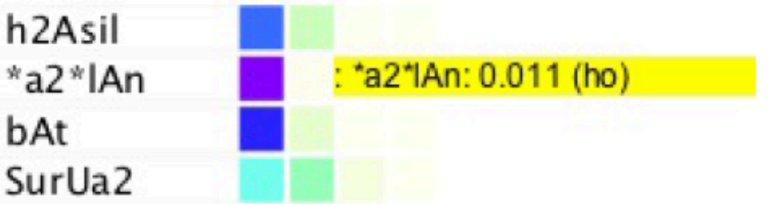
Table 2: Relative frequencies of co-occurrence of nouns with light verbs



Identifying N-V complex predicates in Hindi/Urdu

Visualization tool facilitates...

- zooming and mousing over
  → access to underlying data



- outlier and error detection

V1 in Icelandic (Butt et al. 2014)

V1 (Verb First)

- Common in matrix declaratives in Germanic
- Survived in narrative/joke contexts in other Germanic languages

  Treffen sich zwei Jäger...

  Walked a man into a pub...
- **But:** V1 still common in Icelandic declaratives

Questions:

- What determines the appearance of V1?
- How did this change over time in the history of Icelandic?

V1 in Icelandic

Previous studies (Sigurðsson 1990, Franco 2008, Axel 2009, a.o.) assume that V1 is...

- mainly confined to narrative inversion
- connected to the introduction of unknown referents
- determined by type of subject
- determined by type of verbal element (modals as a marginal phenomenon)

→ a V2 construction with a *pro* or an expletive subject

→ purely syntactic account implies that only unaccusative verbs are possible in V1

V1 in Icelandic

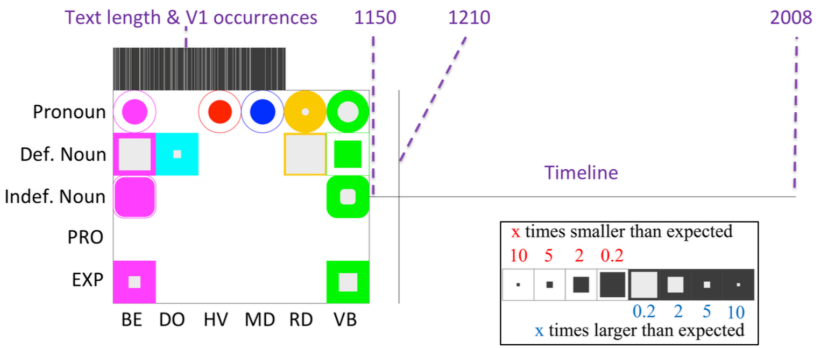
Corpus Study

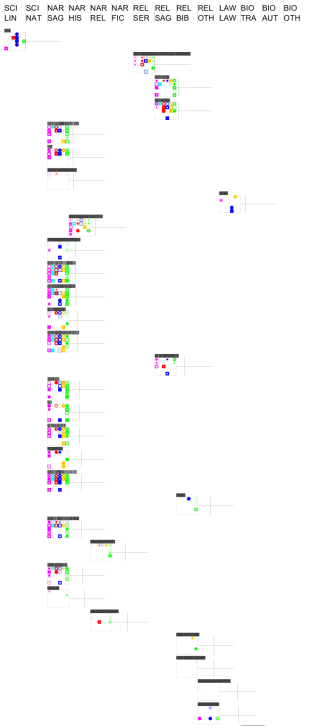
- Icelandic Parsed Historical Corpus (IcePaHC)
  - syntactically annotated (Penn Treebank)
  - 60 texts, 1 million words
  - 12th to 21st century
- Automatic extraction of multiple factors interacting with V1

Visualization

- Glyph visualization allowing for a comparative diachronic evaluation
- Overview First – Details on Demand (Daniel Keim's Mantra)
- Interactive exloratory acces to data

Diachronic Visualization of V1 in Icelandic





Results and Conclusions

V1 declaratives ...

- are not overly associated with subject-less sentences (448 vs. 4893), doubtful that V1 is underlyingly a V2 construction with an empty subject
- do not occur primarily with certain kinds of verbs, not confined to a particular verb class
- mostly found in narrative texts, but occur elsewhere as well.
- There is a marked decrease of V1 as of 1900.

→ Results are much more ‘visible’ compared to results obtained via standard corpus linguistic methods.

→ Visualization plays a key role in furthering the understanding of diachronic data.

→ Tool should be applicable to any diachronic study that seeks to understand a multifactorial interaction.

Evaluation Metrics for Visual Analytics in Linguistics

- Project embedded in the SFB-TRR 161 (University of Stuttgart, University of Konstanz, Max-Planck Institute for Biological Cybernetics Tübingen) funded by the DFG
- **Motivation:**
  - Growing interest in incorporating novel visualization techniques into linguistic research
  - Lack of evaluative methodology and skepticism concerning the value that visual analytics adds to linguistic research
- Comparative evaluation of the usefulness of visual analytics (VA) within linguistics
  - Can we find patterns/insights we could not have found without VA?
  - Can we find patterns/insights more quickly with VA than without?

DFG Deutsche Forschungsgemeinschaft

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

Dative Subjects in Icelandic (Schätzle et al. 2014, Schätzle et al. 2015)

Dative Subjects

- Well-known phenomenon in a multitude of Indo-European languages
- Common in Old-Norse Icelandic and Modern Icelandic

(1) Vel líkuðu goðrøði góð røði.
well like.past.3pl prop.dat.sg good.nom.pl oar.nom.pl
'Goðrøði (the good oarsman) liked good oars well.'
(Fyrsta málfræðiritgerðin, 1150 (IcePaHC))

Debate as to whether...

- dative subjects are a Proto Indo-European Inheritance
- dative subjects are a historical innovation

Dative Subjects in Icelandic

- Indo-Aryan data show evidence for historical innovation (e.g. Butt and Deo 2013, Deo 2003)
- Icelandic has been claimed to show evidence for continuity (e.g. Barðdal et al. 2012)
  $\longrightarrow$ But even there are changes!

Dative Substitution

- Language change in progress: accusatives are original, datives the innovation

(2) Mig langar að fara.
I.acc long.pres to go
'I long to go.'

(3) Mér langar að fara.
I.dat long.pres to go
'I long to go.'

(Smith 1996, 22)

- Smith (1996): datives are systematically associated with either goals or experiencers in modern Icelandic (change in progess)

Dative Subjects in Icelandic

- Complex System of Moving (Diachronic Dimension) and Interacting Parts
  - Case $\longleftrightarrow$ Function (semantics)
  - Grammatical Relations
  - Larger grammar system (division of labor, blocking)
- Some Possible Factors:
  - Verb type/class (verbs with experiencers tend to go dative)
  - Case of objects
  - Word Order
  - Verb voice (active, passive, middle)

Dative Subjects in Icelandic

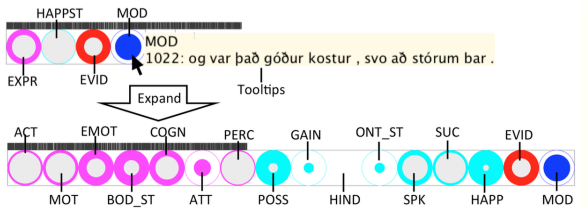
Corpus Study

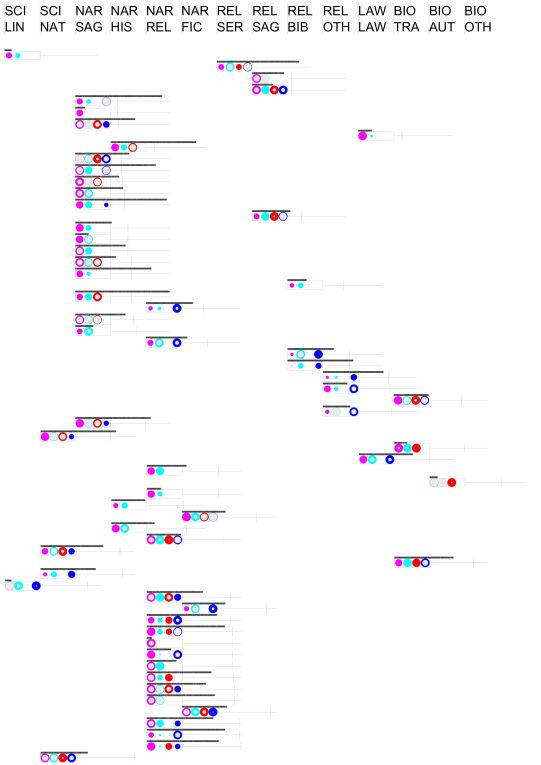
- IcePaHC
- Automatic extraction of verbs inducing dative subjects
- Annotation of lexical semantic verb classes as per the umbrella classification in Barðdal et al. (2012) and the higher classifications for these classes in Barðdal (2011)

Visualization

- Adjustment of V1 visualization (glyph visualization)
- Interaction between verb classes and dative subject occurences
- Overview First – Details on Demand
- Additional expand/fold options

Diachronic Visualization of Dative Subjects in Icelandic





Dative Subjects in Icelandic

Results & Conclusions

- No significant results, but the interactive exploration via novel visualization techniques is on-going
- Preponderance of experience-based and happenstance predicates
- Evidentials and modals appear more often in the latter stage of the language
- Evidence for dative subjects is found in the earliest attested Icelandic texts, but their distribution has been changing over the last millenium
  $\longrightarrow$ Object-to-Subject Hypothesis
- However, the classification system is not ideal for a successful visual analysis...

- ... but there is new data!

Corpus Study II

- Ongoing quantitative analysis of case, voice, word order, and transitivity with focus on dative subjects in matrix declarative sentences in IcePaHC.
- Different verb types given by the annotation scheme of the corpus are analyzed separately
  - main verbs
  - modals
  - have
  - future work: do, become, be
- Division of data into time stages suggested for Icelandic (Haugen 1984)
- χ^2-test in order to show whether the observed distributions differ from the expected distributions (* $p<0.05$, ** $p<0.01$, *** $p<0.001$).

Corpus Study II – Main Verbs, Datives and Middles

Some Results

- For all verb types, dative subjects are on the increase.
- This increase is strongly correlated with a rise in middle forms $\longrightarrow$ closer look:

Time	active	middle	passive	χ^2
until 1350	70.4%	16.6%	13.0%	***
until 1550	69.3%	16.4%	14.3%	***
until 1750	49.2%	21.8%	29.0%	***
until 1900	58.5%	24.3%	17.2%	
until 2008	41.0%	47.1%	11.8%	***
all	58.0%	25.9%	16.1%	

Table: Dative subjects with main verbs according to voice in %

- The use of dative subjects with middles increases over time.
- Active constructions appear less often with dative subjects.

Corpus Study II - Middle voice

```
(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))
```

(4) Mér finnst sem ég sé sloppinn úr einhvers
l.dat.sg seem.pres.mid.sg like l.nom.sg be.subjunc.pres.1.sg escape.ppart.m.nom.sg from some.gen.sg
konar fangelsi.
kind.gen.sg prison.dat.sg
'It seems to me as if I have escaped from some kind of prison.'
(Brynjólfur Sveinsson biskup, 1882)

Corpus Study II - Summary of Results

- Overall, dative subjects are on the increase.
- There is a strong association with middles.
- These middles are mainly found on:
  - Psych Predicates (experiencer verbs)
  - ‘seem’ (raising predicates)
- Nominative subjects are correspondingly decreasing.
- Other possible factors (e.g., word order, modality, verb type) do not have a significant impact.

→ Systematic innovation of experiencer predicates via middle formation

→ Dative case is becoming more systematically associated with lexical semantic factors

→ Complex overall system with many moving parts

Outlook

Future work

- Experiment with a different classification of verbs, broader umbrella classes
- Further Exploration of Possibilities offered by Visual Analytics
  - The systems illustrated here are very new.
  - Interactive exploratory linguistic analysis is on-going.
  - Systems are being fine-tuned.

Workflow

- Use cases for Digital Humanities/eHumanities are being developed.
- Infrastructure Platforms (mix and match the available tools)

Measuring Success

- Development of Evaluation Metrics for LingVis.
- Use cases, workflow and result comparison.

Thank you!
Questions?

Linguistics

Tina Bögel, Miriam Butt, Annette Hautli-Janicz, Thomas Mayer, Maike Müller, Frans Plank, Christin Schätzle

Computer and Information Sciences

Daniel Keim, Menna El-Assady, Andreas Lamprecht, Christian Rohrdantz, Dominik Sacha

Political Science

Katharina Holzinger, Valentin Gold

Some References

Butt, Miriam. 2015. Visualizing Linguistic Structure (LingVis). Talk given at the AAAS Annual Meeting as part of the Symposium on *Visualizing Verbal Culture: Seeing Language Diversity*.

Butt, Miriam, Bögel, Tina, Kotcheva, Kristina, Schätzle, Christin, Rohrdantz, Christian, Sacha, Dominik, Dehe, Nicole and Keim, Daniel. 2014. V1 in Icelandic: A multifactorical visualization of historical data. In *Proceedings of VisLR: Visualization as added value in the development, use and evaluation of Language Resources*, Workshop at the 9th edition of the Language Resources and Evaluation Confernece (LREC 2014), Reykjavik, Iceland.

Butt, Miriam and Deo, Ashwini. 2013. A Historical Perspective on Dative Subjects in Indo-Aryan, paper presented at the *LFG13 Conference*.

Butt, Miriam, Tina Bögel, Annette Hautli, Sebastian Sulger and Tafseer Ahmed. 2012. Identifying Urdu Complex Predication via Bigram Extraction. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*, 409–424. Mumbai, India.

Deo, Ashwini. 2003. Valency Change and Case Marking: Marathi Dative Experiencers, handout from the *Pioneer Workshop on Case, Valency and Transitivity*.

Hautli-Janisz, Annette. 2015. Deliberation in Political Negotiation: a Linguistically-motivated Visual Analysis. Talk in the *ARG-tech group*, University of Dundee.

Rohrdantz, Christian, Michael Hund, Thomas Mayer, Benjamin Wälchli, and Daniel A. Keim. 2012. The World's Languages Explorer: Visual Analysis of Language Features in Genealogical and Areal Contexts. *Comput. Graph. Forum 31 (3)*, 935–944.

Gold Valentin, Mennatallah El-Assady, Tina Bögel, Christian Rohrdantz, Miriam Butt, Katharina Holzinger and Daniel Keim. 2015. Visual Linguistic Analysis of Political Discussions: Measuring Deliberative Quality. *Digital Scholarship in the Humanities 2015*.

Schätzle, Christin, Kristina Kotcheva and Miriam Butt. 2015. The Diachronic Development of Dative Subjects in Icelandic, paper presented at the *LFG15 Conference*.

Schätzle, Christin, Sacha, Domnik and Butt, Miriam. 2014. Diachronic Visualization of Dative Subjects in Icelandic, poster presentation at the Workshop on *Big Data Visual Computing* at the 44th Annual Meeting of the Gesellschaft für Informatik in Stuttgart, 22nd September 2014.