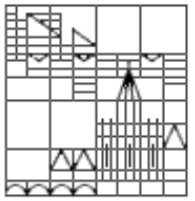
SPONSORED BY THE
Federal Ministry
of Education
and Research



Visual Analytics for Linguists

Miriam Butt & Chris Culy

ESSLII 2014, Introductory Course

Tübingen

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

PHILOSOPHISCHE
FAKULTÄT
Seminar für Sprachwissenschaft

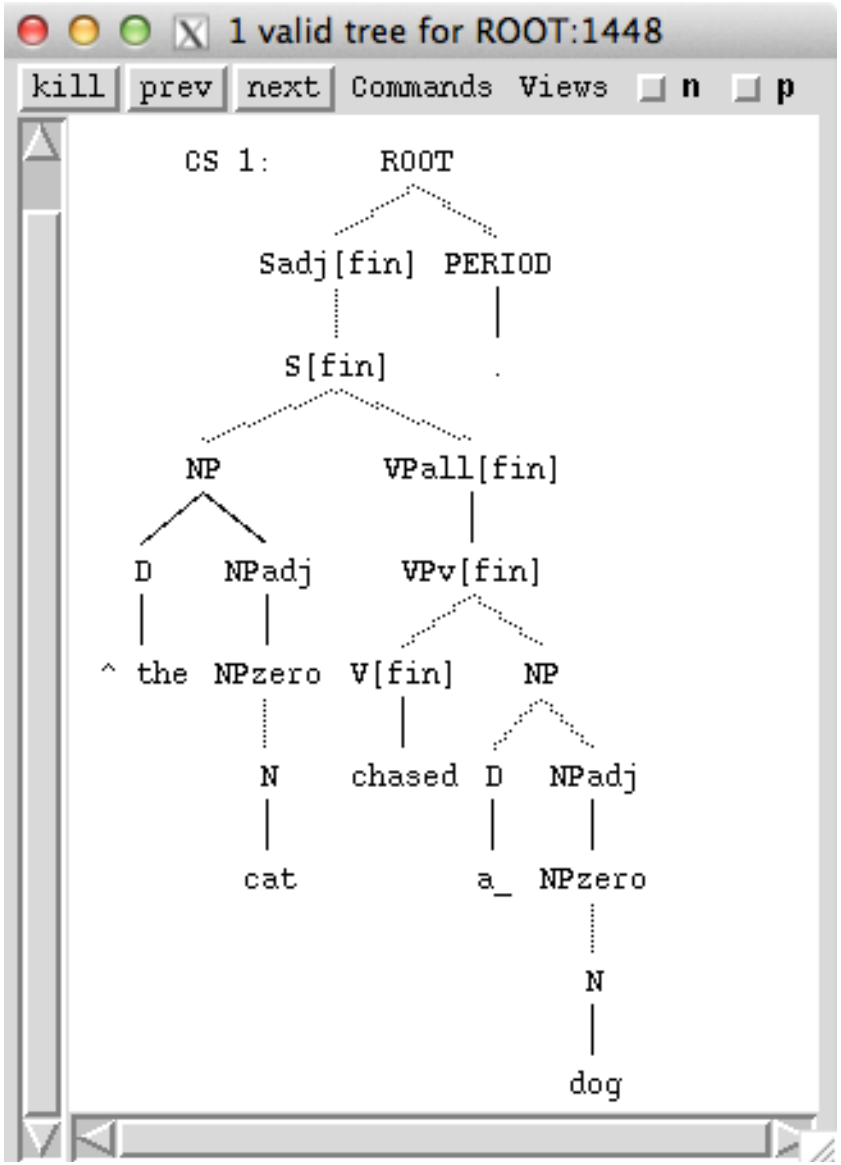
Day 3 – Towards Hands On

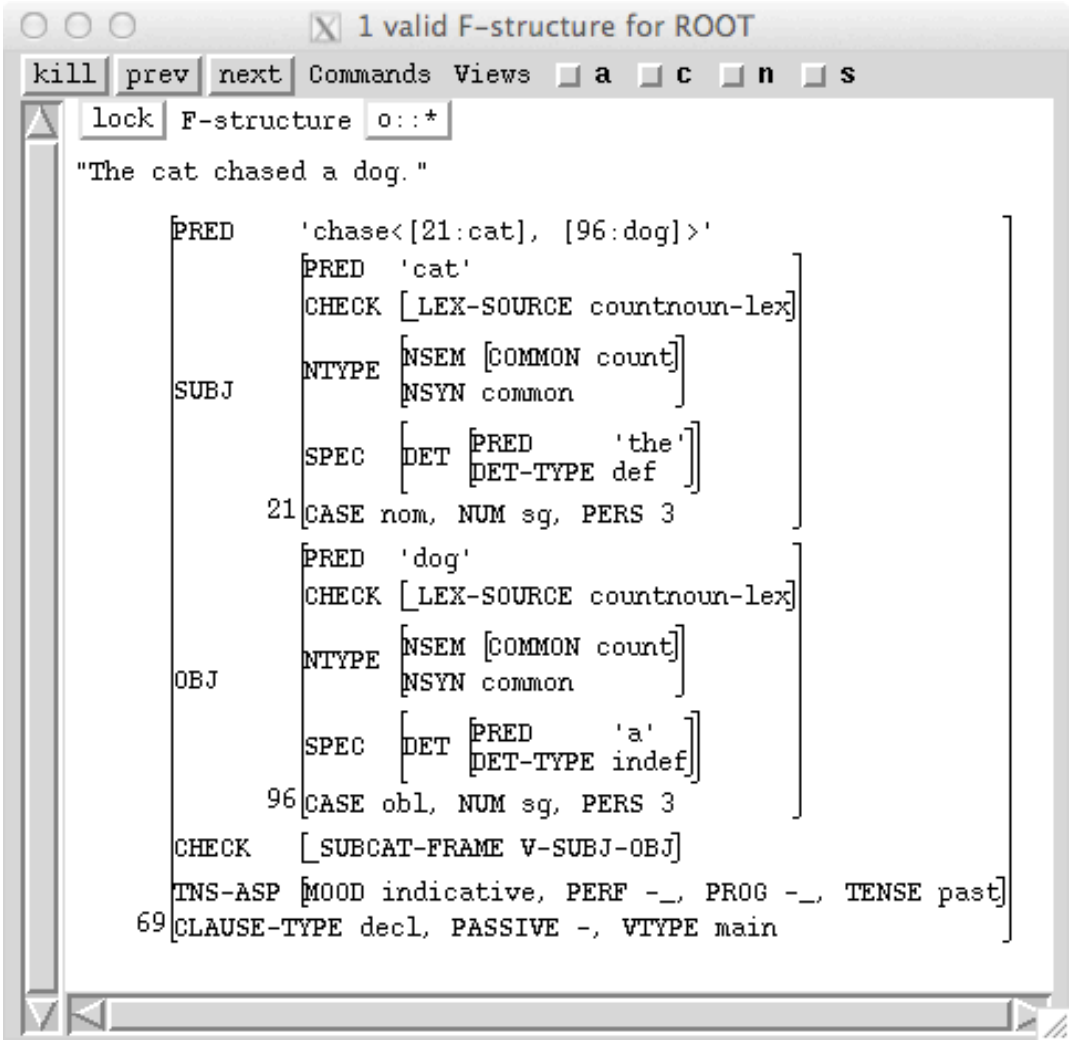
1. More Use Cases and Critical Discussion
   - Are the visualizations successful?
   - Is the linguistic content modeled insightfully?
2. Presentation of Hands-On Visualizations

Tree Comparison via Sunburst

- Various types of trees are used to represent data in terms of hierarchical relationships.
  - XML hierarchies
  - Linguistic structure
- Concrete Example: LFG
  - c-structures via standard trees
  - f-structures: dependency structure via AVMs

English LFG ParGram Grammar

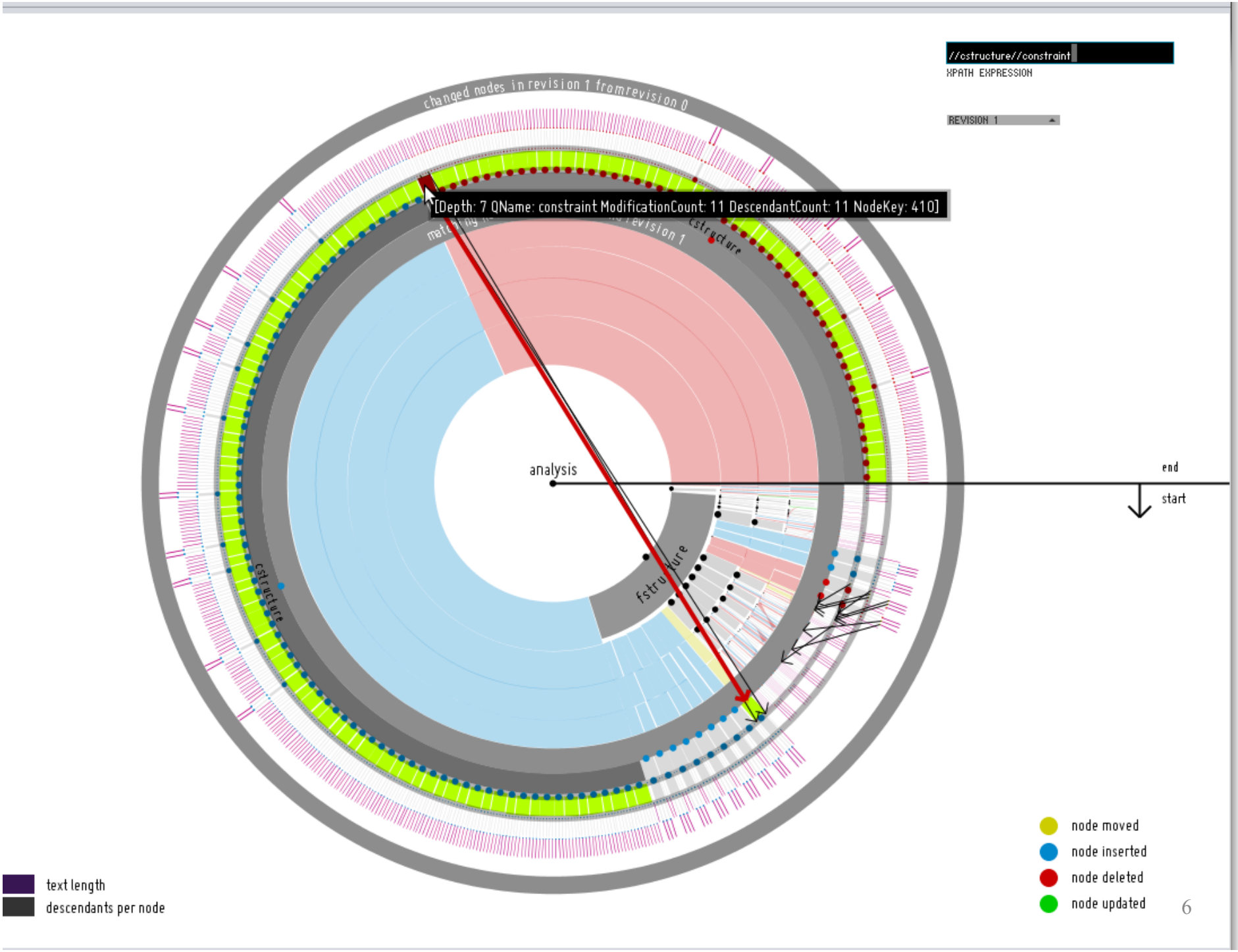


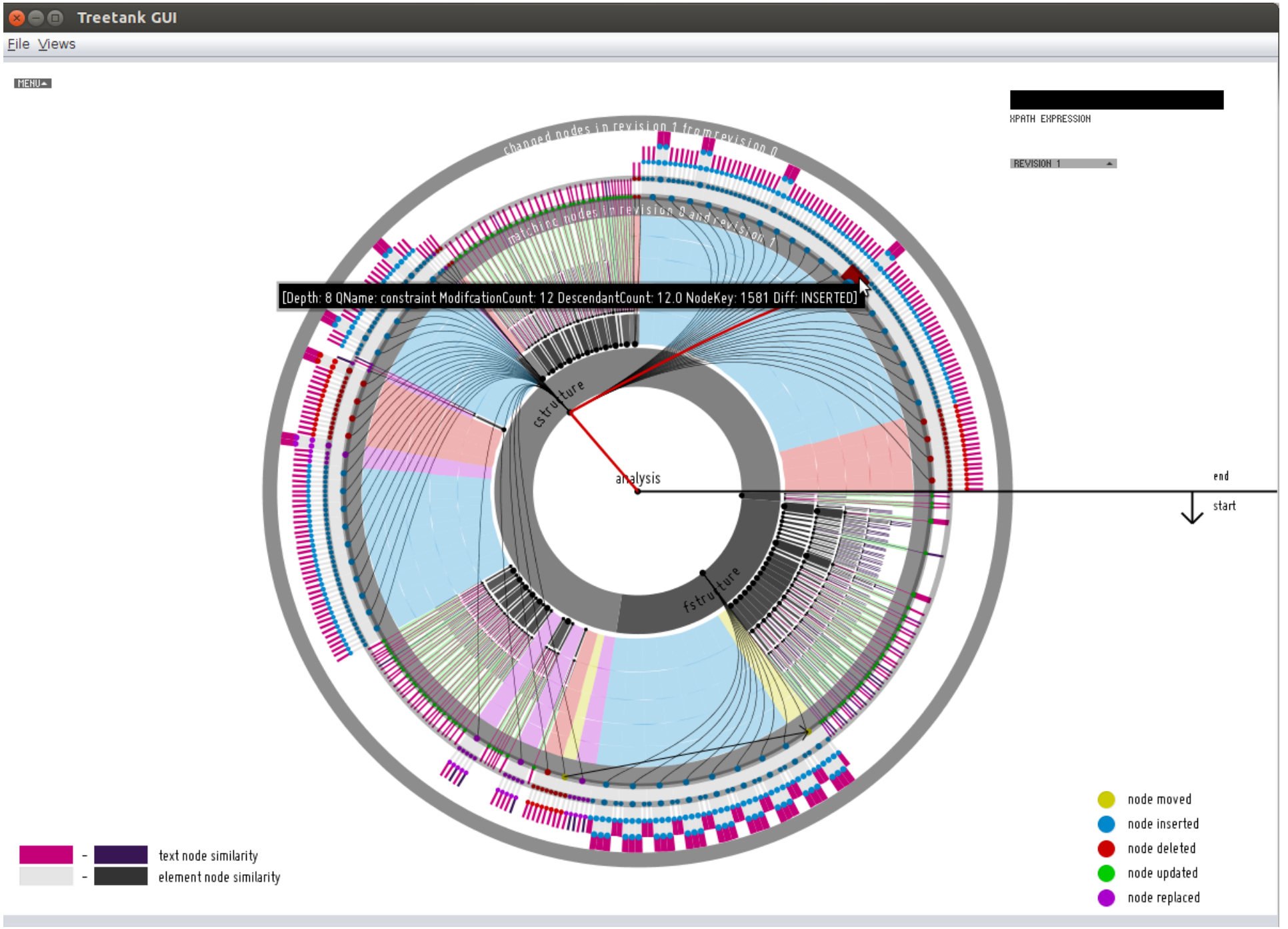


Tree Comparison for Grammar Development

- In Grammar Development the grammar is routinely updated/changed.
- This necessarily means that the output will differ.
- Would be good to have an automatic visual tree comparison method.
- The following are proposals by Lichtenberger(2012).

Johannes Lichtenberger. 2012. A Visual Analytics Approach for Comparing Tree Structures. MSc., University of Konstanz. http://kops.ub.uni-konstanz.de/handle/urn:nbn:de:bsz:352-210419


//cstructure//constraint
XPATH EXPRESSION
REVISION 1
changed nodes in revision 1 fromrevision 0
[Depth: 7 QName: constraint ModificationCount: 11 DescendantCount: 11 NodeKey: 410]
cstructure
analysis
end
start
fstructure
cstructure
node moved
node inserted
node deleted
node updated
text length
descendants per node

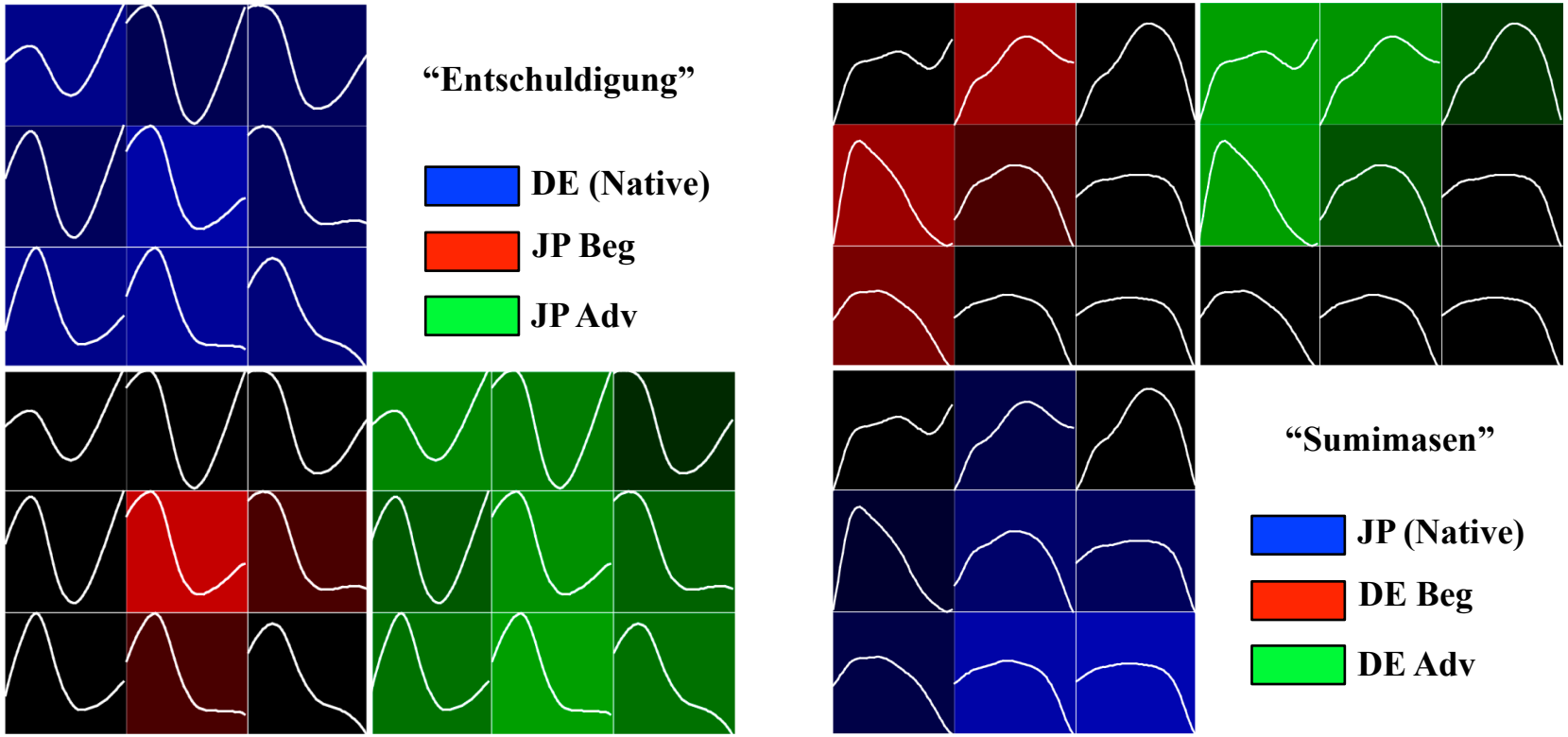

Treetank GUI
File Views
MENU
XPATH EXPRESSION
REVISION 1
changed nodes in revision 1 from revision 0
matching nodes in revision 0 and revision 1
[Depth: 8 QName: constraint ModifcationCount: 12 DescendantCount: 12.0 NodeKey: 1581 Diff: INSERTED]
cstructure
analysis
fstructure
end
start
node moved
node inserted
node deleted
node updated
node replaced
text node similarity
element node similarity

Visualization of Pitch Contours

- So far we have been working with textual data.
- However, one can also work with spoken data.
- For Visual Analytics, all one needs is to have features (or vectors) that can be computed with.

- **Example:**
  - Analysis of Pitch Contours via Self-Organizing Maps
  - in combination with Visual Analytics
- **Data**
  - Japanese vs. German ‘sorry’
  - Japanese pitch contour always has a fall
  - Germans can vary according to pragmatic intent
  - Recorded German and Japanese natives
  - vs. learners of German and Japanese (beginners/ advanced)
  - learners of Japanese were German and vice versa

German Entschuldigung ‘sorry’ vs. Japanese Sumimasen ‘sorry’



Self-Organizing Maps

Visualization Demo

Cluster Visualization

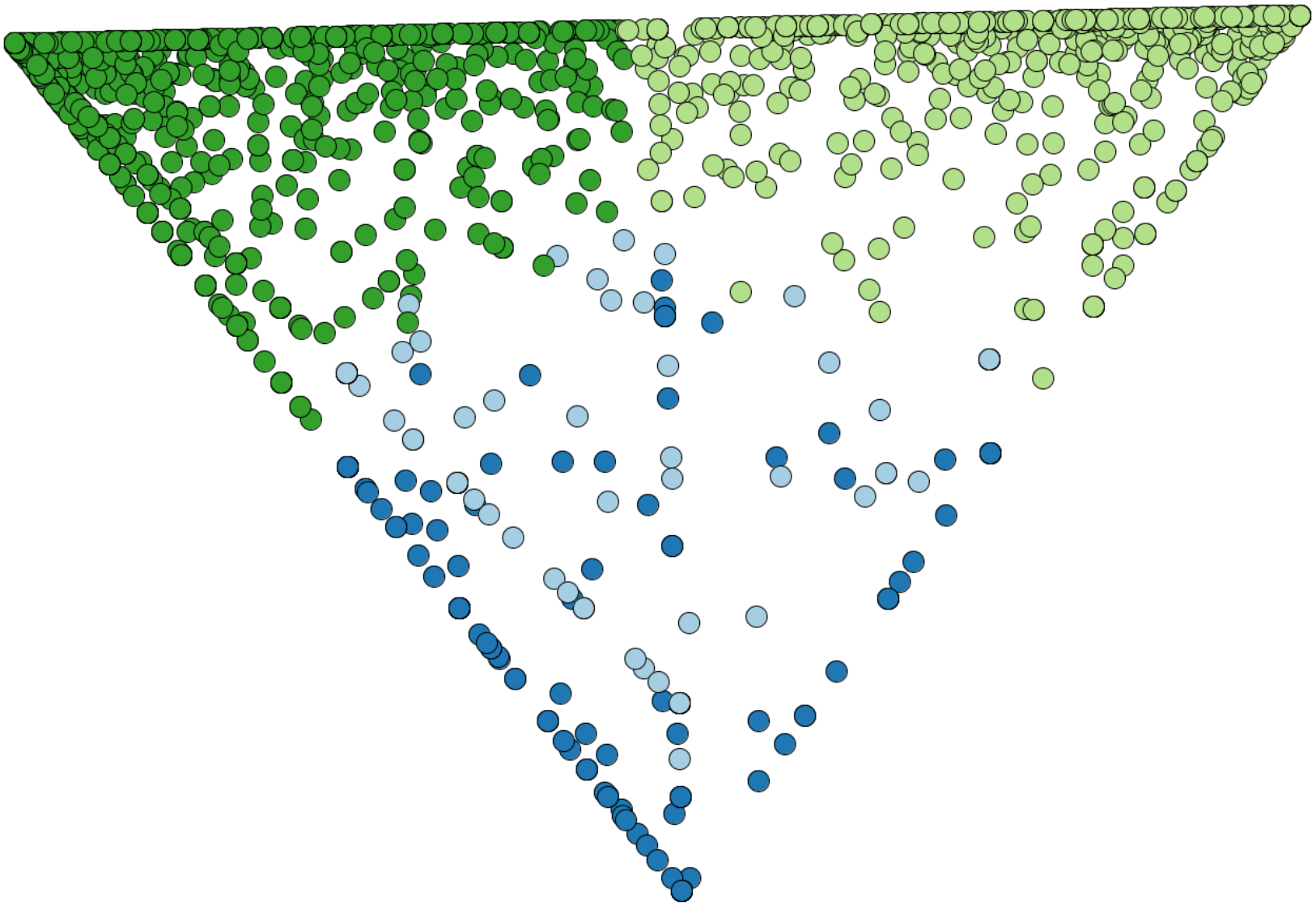
- Automatic clustering methods are increasingly being used by a wide range of linguists.
- However, it is often hard to understand what the clustering method is doing.
- And it is hard to interact with it.
- The following presents an interactive, flexible visual analytic approach to clustering information.

Lamprecht, Andreas, Hautli, Annette, Rohrdantz, Christian and Tina Bögel. 2013. 'A Visual Analytics System for Cluster Exploration'. In *Proceedings of ACL 2014* , 109-114, Sofia, Bulgaria

Cluster Visualization

- So far allows for standard k-means or GVM clustering.
- **Important Note**: the visualization adds the visual and interactive component – it does not improve on the statistical approaches per se.
- Each data point is represented by a dot.
- The user can specify the amount of clusters desired.

Sample Visualization

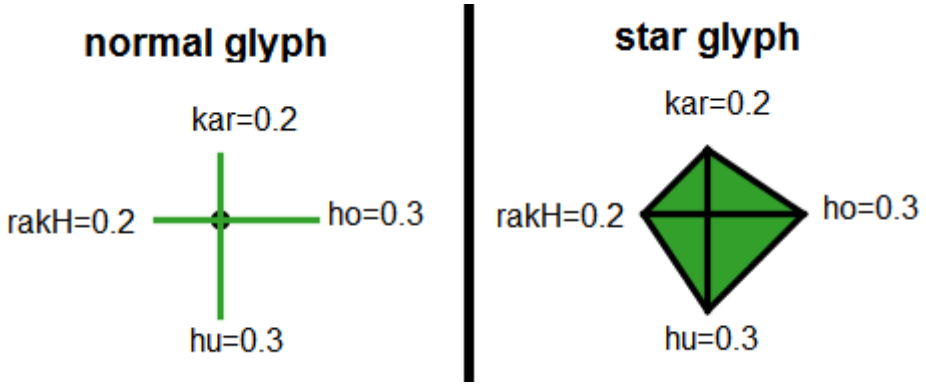


Introducing Glyphs

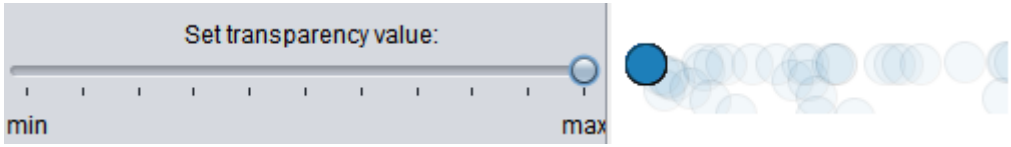
- Glyphs are combinations of symbols that are defined to have a certain meaning.
- Data objects in the visualization can be presented either as circles, normal glyphs or star glyphs.
  - Circles: Every noun represented by a colored circle
  - Normal glyphs: Relative bigram frequencies mapped onto the length of arcs (ordered clock-wise around the center beginning in north position)
  - Star glyphs: Extension of normal glyphs, ends of arcs are connected to form a “star".

Introducing Glyphs

- The data shown here is that from the N-V complex predicates of Day 1.
- There are four light verbs (*kar* ‘do’, *ho* ‘be’, *hu* ‘become’ and *rakh* ‘put’).
- The numbers show the frequency with which they appear with a given noun – the data point represented by the dot.

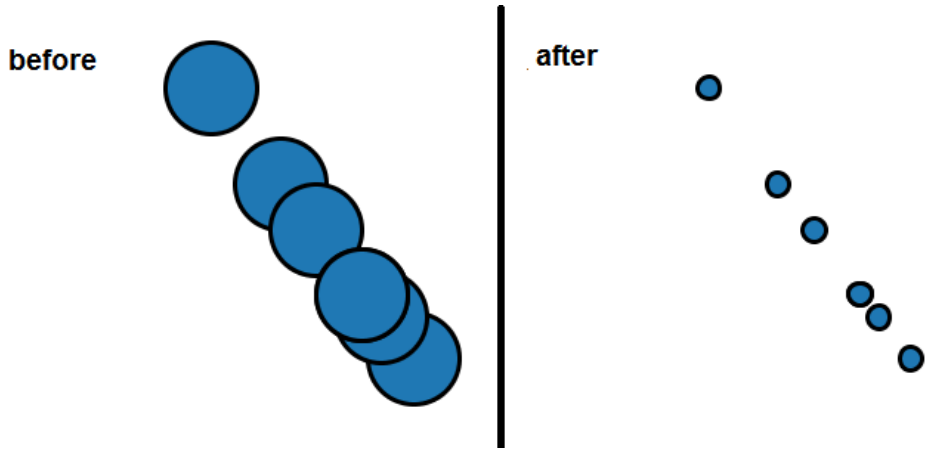


- Overplotting is a problem when the data set becomes large or when the data points are very similar to one another.
- Several strategies to handle this interactively:
  - change transparency of objects



  - reposition data objects

  - scale data objects



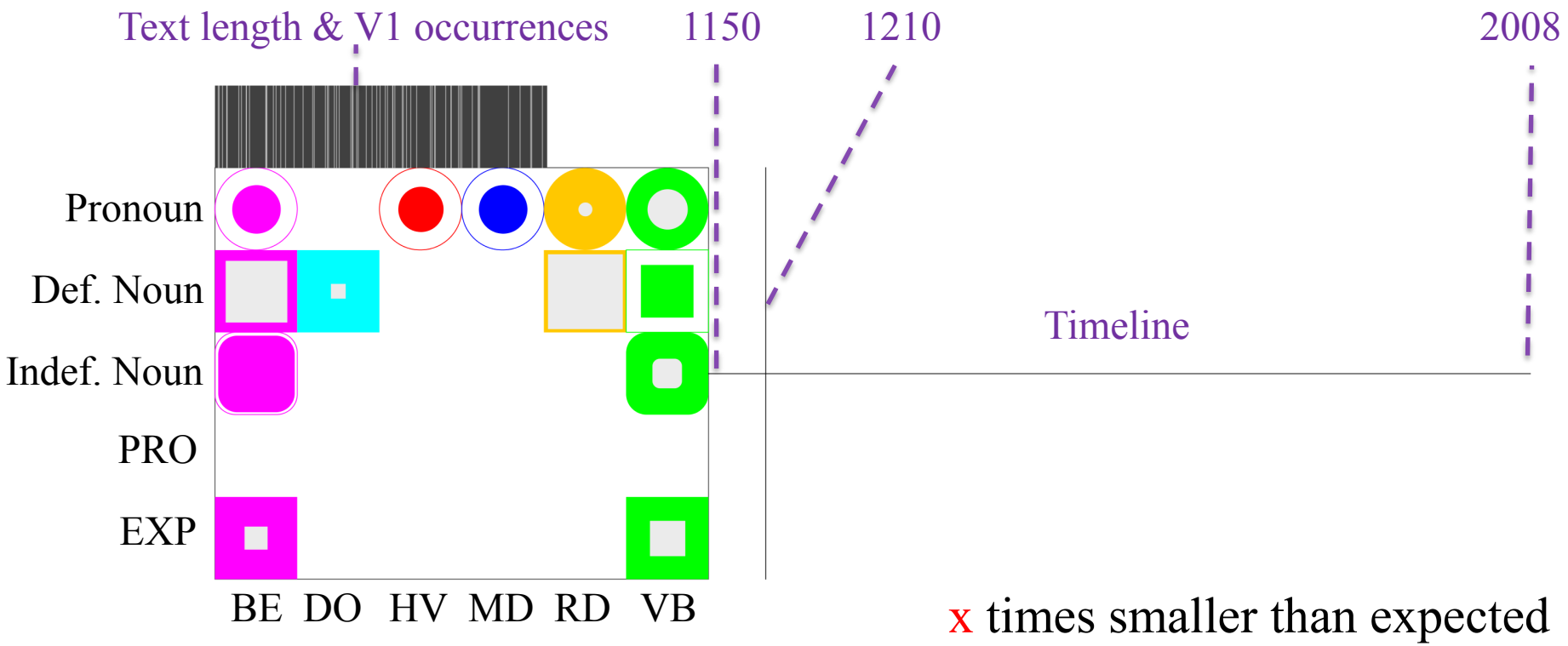
Improved Cluster Visualization Demo

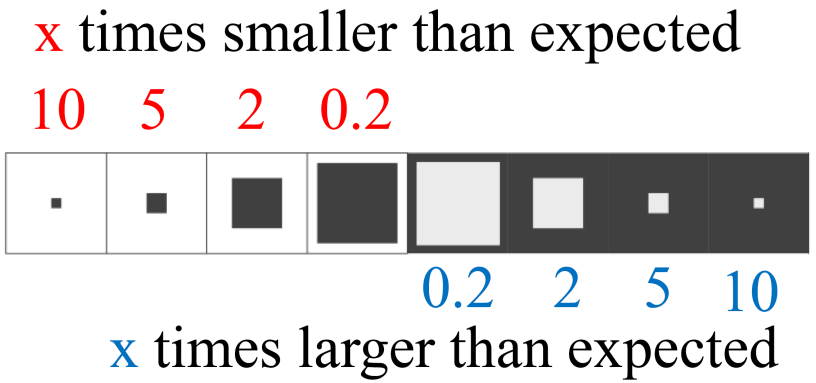
More Glyphs – Icelandic Data

- We have recently been working with the annotated diachronic corpus of Icelandic (IcePaHC).
- Two Questions so far:
  - When does V1 in Icelandic Occur?
  - What governs the appearance of dative subjects.
- Both of these questions have been of great interest for linguists.

Butt, Miriam, Bögel, Tina, Kotcheva, Kristina, Schätzle, Christin, Rohrdantz, Christian, Sacha, Dominik, Dehé, Nicole and Daniel Keim. 2014. 'V1 in Icelandic: A Multifactorical Visualization of Historical Data'. *Proceedings of the LREC 2014 Workshop on Visualization as added value in the development, use and evaluation of LRs (VisLR).* Reykjavik, Iceland

Factors identified by linguists as being relevant to V1 in Icelandic





- The entire Icelandic corpus.
- Patterns become apparent quite quickly.
- Can zoom in and investigate in more detail.

Icelandic Visualization Demo