Hands-On Session

Miriam Butt & Dominik Sacha

LingVis: Visual Analytics for Linguistics
DGfS 2016 | 24.-26.2.2016

Hands-On

- **We can work with:**
  - Self Organizing Maps for Pitch Contour Analysis
  - Diachronic Visualization of Language Properties
  - Cluster Visualization
  - The WALS Explorer
  - DoubleTreeJS/KWIC

- **In the following, we explain:**
  - how to work with them
  - type of data needed

- **Also pointers to further sites**

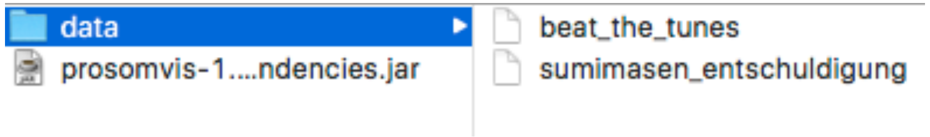
Self Organizing Maps for Pitch Contour Analysis

Analyzing Spoken Data

D. Sacha, Y. Asano, C. Rohrdantz, F. Hamborg, D. A. Keim, B. Braun and M. Butt. Self Organizing Maps for the Visual Analysis of Pitch Contours. Proceedings of the 20th Nordic Conference of Computational Linguistics. Vilnius, Lithuania, ACL Anthology, 23():181-190, 2015.

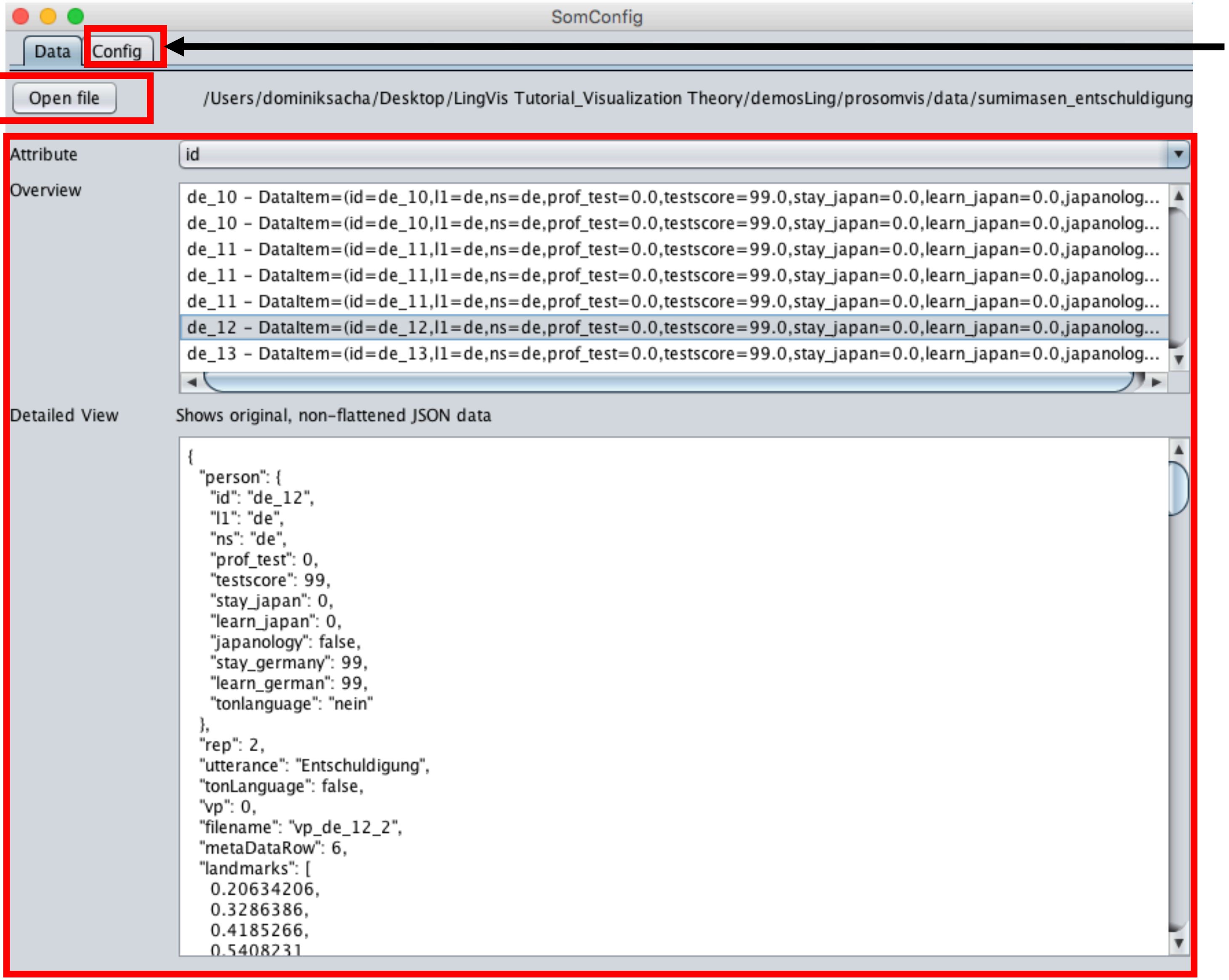
SOM – Starting the Tool

- All material (software and data) is contained in the “prosomvis” folder



- Double-Click the prosomvis-Jar file
  - prosomvis-1.0.0-SNAPSHOT-jar-with-dependencies.jar
  - The App should start

SOM– Loading Data



/Users/dominiksacha/Desktop/LingVis Tutorial_Visualization Theory/demosLing/prosomvis/data/sumimasen_entschuldigung

Attribute: id

Overview

```
de_10 - DataItem=(id=de_10,l1=de,ns=de,prof_test=0.0,testscore=99.0,stay_japan=0.0,learn_japan=0.0,japanolog...
de_10 - DataItem=(id=de_10,l1=de,ns=de,prof_test=0.0,testscore=99.0,stay_japan=0.0,learn_japan=0.0,japanolog...
de_11 - DataItem=(id=de_11,l1=de,ns=de,prof_test=0.0,testscore=99.0,stay_japan=0.0,learn_japan=0.0,japanolog...
de_11 - DataItem=(id=de_11,l1=de,ns=de,prof_test=0.0,testscore=99.0,stay_japan=0.0,learn_japan=0.0,japanolog...
de_11 - DataItem=(id=de_11,l1=de,ns=de,prof_test=0.0,testscore=99.0,stay_japan=0.0,learn_japan=0.0,japanolog...
de_12 - DataItem=(id=de_12,l1=de,ns=de,prof_test=0.0,testscore=99.0,stay_japan=0.0,learn_japan=0.0,japanolog...
de_13 - DataItem=(id=de_13,l1=de,ns=de,prof_test=0.0,testscore=99.0,stay_japan=0.0,learn_japan=0.0,japanolog...
```

Detailed View: Shows original, non-flattened JSON data

```
{
 "person": {
  "id": "de_12",
  "l1": "de",
  "ns": "de",
  "prof_test": 0,
  "testscore": 99,
  "stay_japan": 0,
  "learn_japan": 0,
  "japanology": false,
  "stay_germany": 99,
  "learn_german": 99,
  "tonlanguage": "nein"
 },
 "rep": 2,
 "utterance": "Entschuldigung",
 "tonLanguage": false,
 "vp": 0,
 "filename": "vp_de_12_2",
 "metaDataRow": 6,
 "landmarks": [
  0.20634206,
  0.3286386,
  0.4185266,
  0.5408231
```

1. Click on the Button and load a data file

2. You may inspect the loaded data

3. Finally switch to the config tab

SOM– Configuration (Simple)

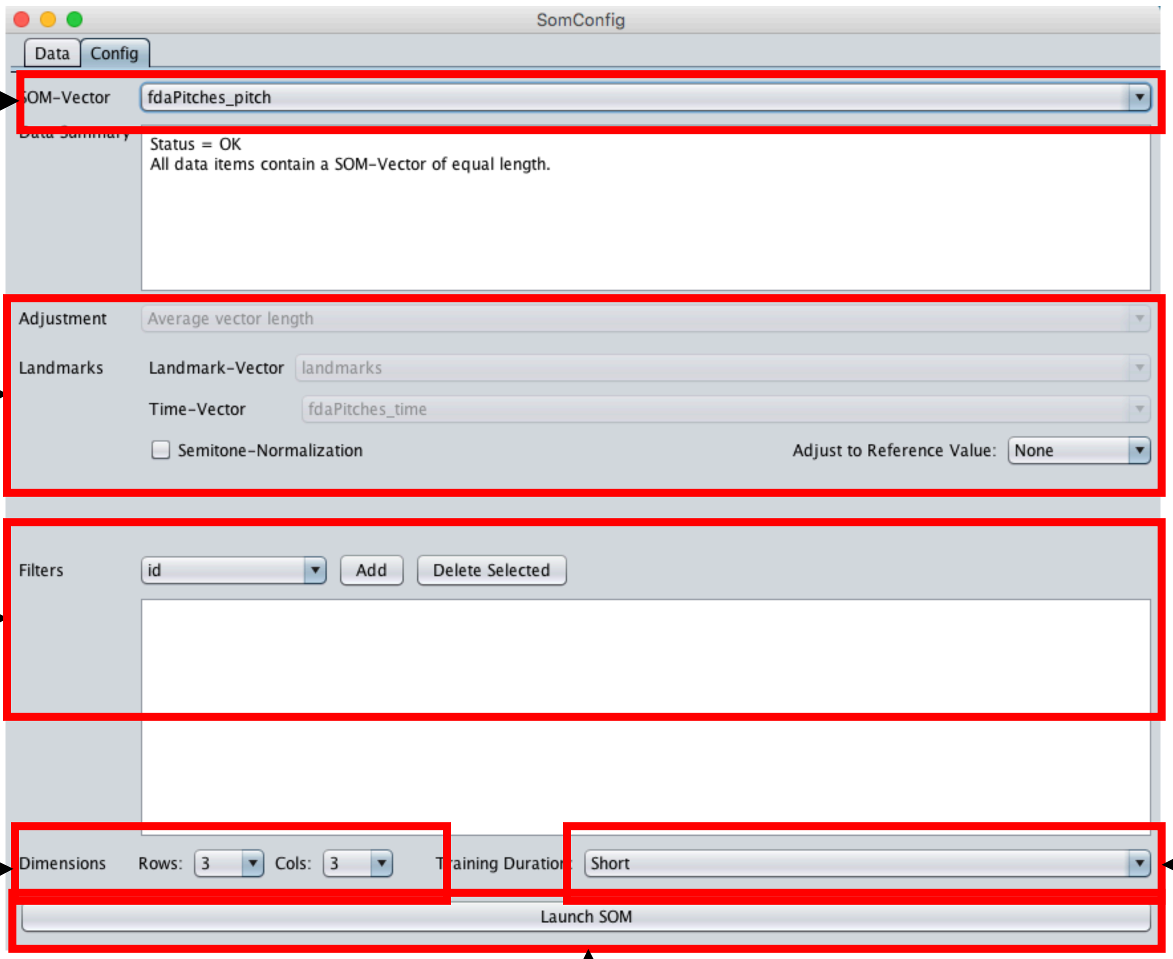
1. Select the Vector that shall be analyzed

(For the sumimasen dataset a preprocessed vector is included)

2. If no preprocessed vector is detected, further preprocessings have to be applied

3. A data filter may be applied e.g., utterance="sumimasen" Note, data may be filtered during the analysis process as well

4. Set the SOM dimensions. Feel free to try different ones



SOM-Vector: fdaPitches_pitch

Data Summary:
Status = OK
All data items contain a SOM-Vector of equal length.

Adjustment: Average vector length

Landmarks: Landmark-Vector: landmarks

Time-Vector: fdaPitches_time

☐ Semitone-Normalization

Adjust to Reference Value: None

Filters: id

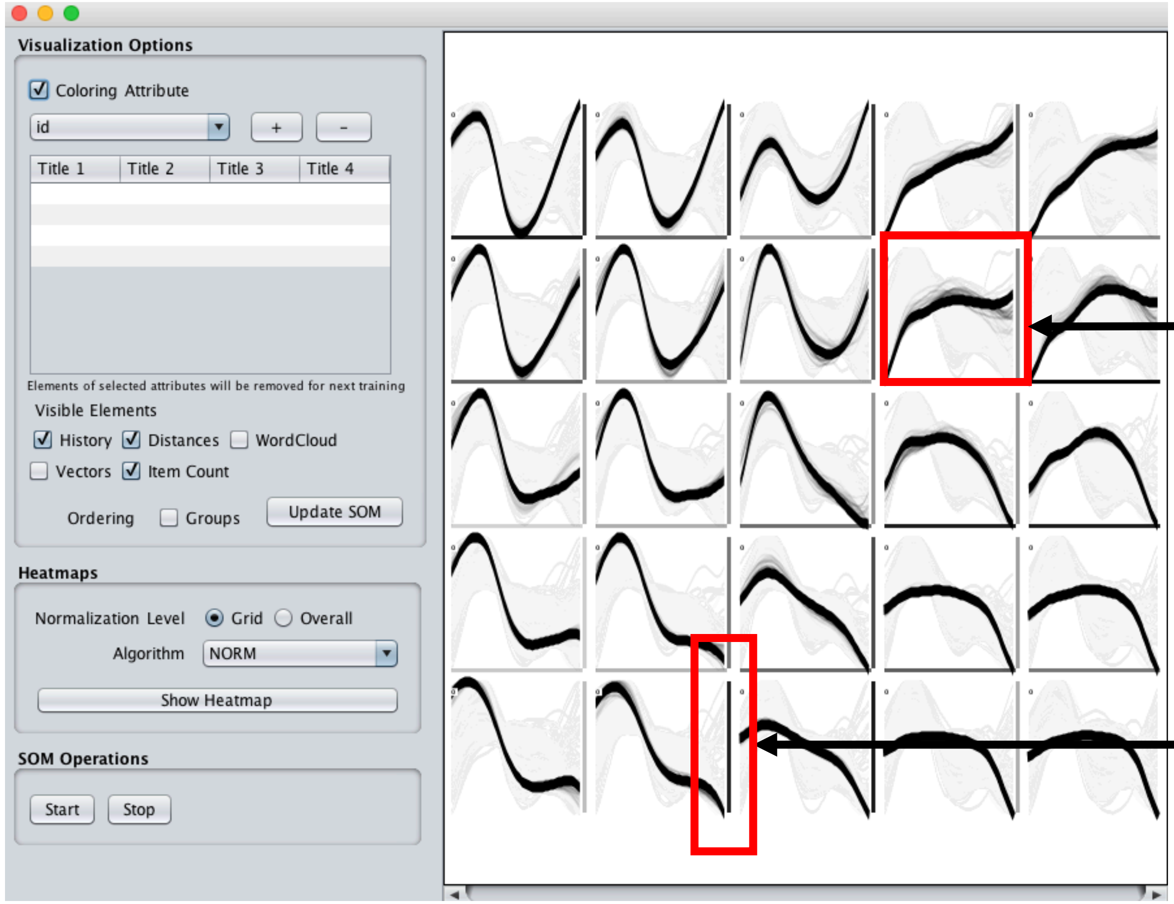
Dimensions Rows: 3 Cols: 3

Training Duration: Short

5. The SOM training may be "long" or "short" Suggestion: Try short first and switch to a longer training to confirm a concrete hypothesis

6. Finally launch the SOM

SOM Training – Just Watch



Visualization Options

☑ Coloring Attribute

id

Title 1	Title 2	Title 3	Title 4

Elements of selected attributes will be removed for next training

Visible Elements

☑ History ☑ Distances □ WordCloud

□ Vectors ☑ Item Count

Ordering □ Groups

Heatmaps

Normalization Level ● Grid ○ Overall

Algorithm NORM

SOM Operations

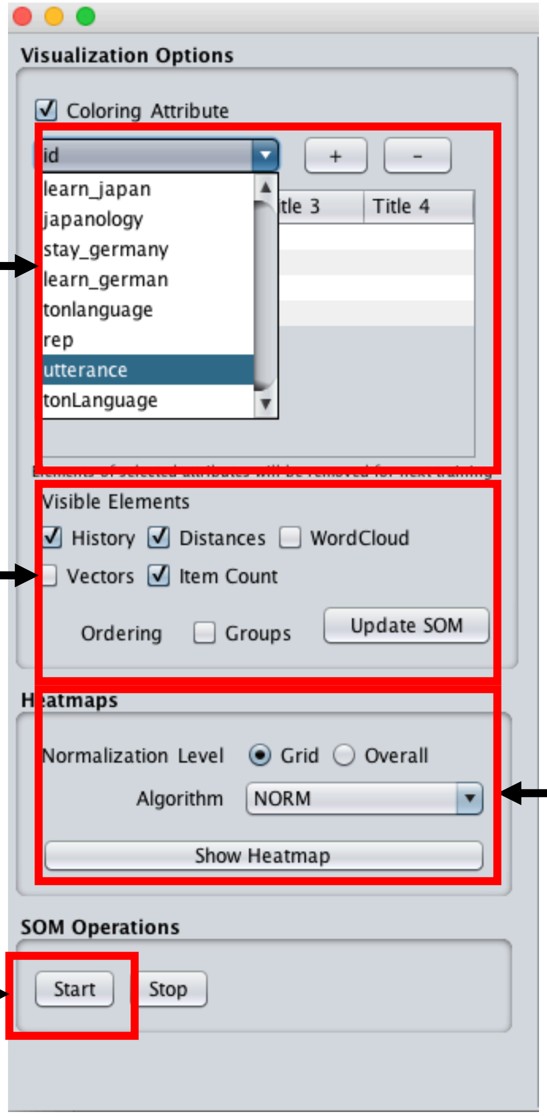
Observe the changing cells (contours, background)

Observe the distance bars

SOM Visualization - Configuration

1.) Add/remove attributes to be visualized. E.g., “utterance” or “ns” (Hint: Start with only one attribute)

2.) Try out different grid and cell visualizations (if the checkboxes do not work use the update button)



☑ Coloring Attribute

Visible Elements

☑ History ☑ Distances □ WordCloud

□ Vectors ☑ Item Count

Ordering □ Groups

Normalization Level ● Grid ○ Overall

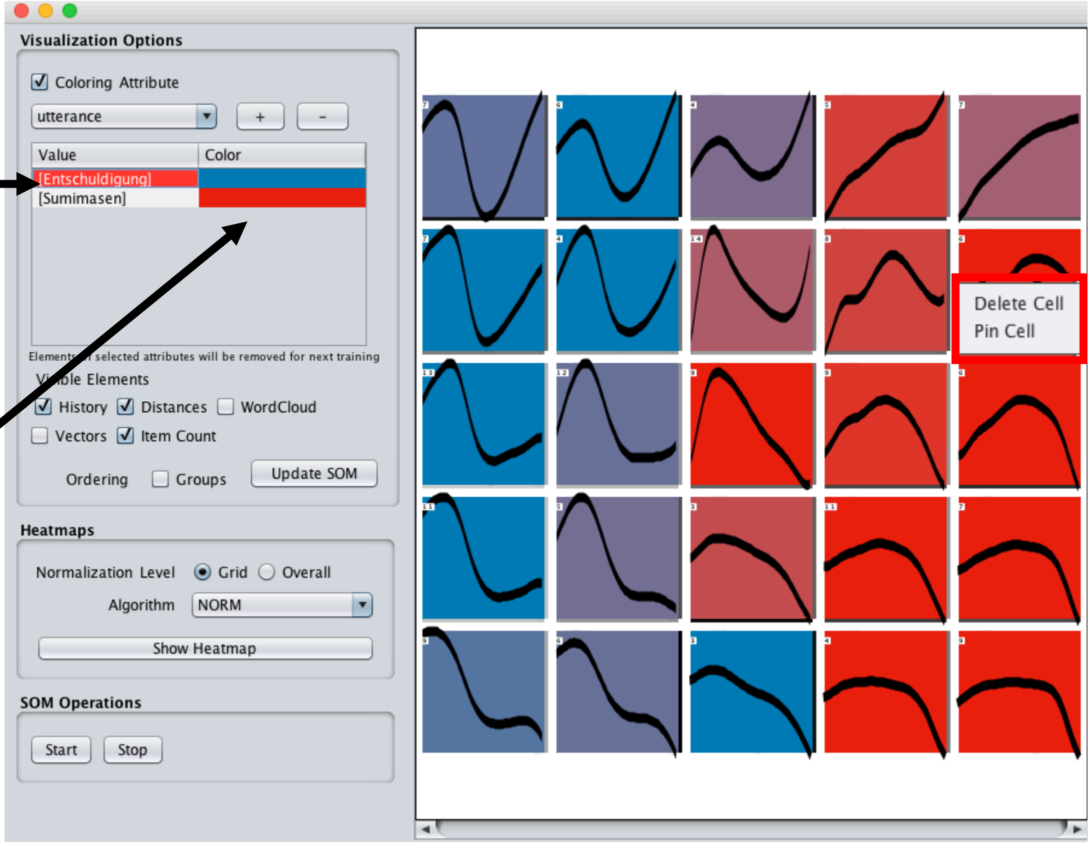
Algorithm NORM

3.) Once an attribute has been selected (1.)), a separate heatmap may be opened for each attribute value. Try out different normalization techniques

4.) A subsequent SOM may be trained with the remaining data items

SOM Visualization – Data Filtering & Grid Interaction

1.) Select an attribute to exclude it for the next SOM training



Visualization Options

☑ Coloring Attribute

utterance ▾ | + | -

Value	Color
[Entschuldigung]	🟦
[Sumimasen]	🟥

Elements of selected attributes will be removed for next training

Visible Elements

☑ History ☑ Distances □ WordCloud

□ Vectors ☑ Item Count

Ordering □ Groups

Heatmaps

Normalization Level ● Grid ○ Overall

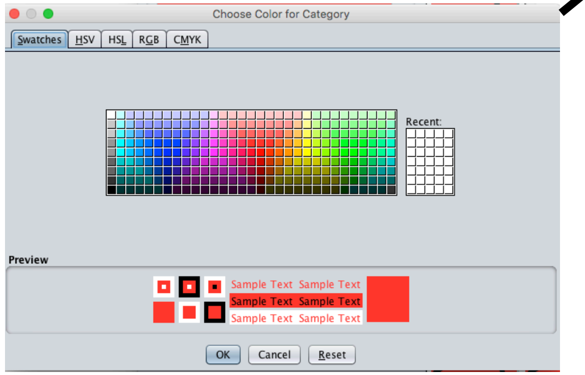
Algorithm NORM ▾

SOM Operations

2.) Right click on a cell to open the contect menue (cells may be deleted or pinned)

3.) Cells may be drag & dropped (moved) & pinned

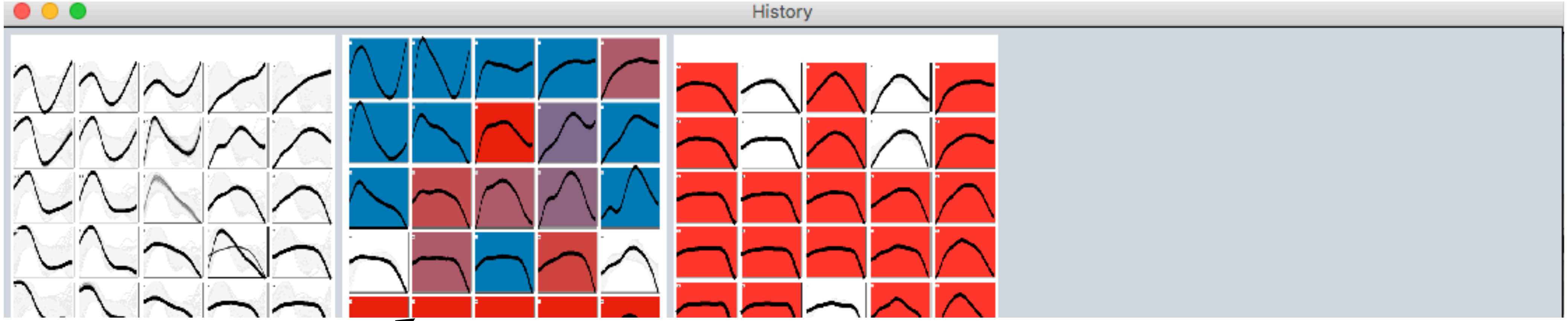
4.) Choose another color (click on the Color cell)



Choose Color for Category

Preview

SOM - History



1.) Click on the SOMs to bring them to the front of the screen

Some Hints - I

- Sumimasen_Entschuldigung
  - Data set contains recorded utterances “sumimasen” (JP) & “entschuldigung” (GER) [--> Excuse me]
  - German and Japanese speakers and learners (L1, L2)
  - Dataset is well preprocessed and needs no further configuration

  - Visualize the whole dataset (select the utterance attribute)
  - Filter out the “Entschuldigung” Utterances (in the 1st SOM)
  - Visualize the different speaker nationalities (”ns” attribute)

Sumimasen - Example

- **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

Some Hints

- Beat the tunes
  - Data set of recorded non-sense words of Japanese and German speakers
  - A lot of meta-data is contained
  - Data set is not preprocessed

  - Try out different preprocessings
  - Visualize the whole dataset (select the “pitch” attribute)
  - Filter out the “HH” Utterances (in the 1st SOM)
  - Train another SOM (based on the 1st SOM) that filters out the “HL”
  - Analyze the different dialects/segments/speakers (use the heatmaps)

Diachronic Visualization of Text Properties/Icelandic

Analyzing Language Change Based on Feature Occurrences

M. Butt, T. Bögel, K. Kotcheva, C. Schätzle, C. Rohrdantz, D. Sacha, N. Dehe and D. A. Keim. V1 in Icelandic: A multifactorical visualization of historical data. Proceedings of the LREC 2014 Workshop VisLR: Visualization as added value in the development, use and evaluation of Language Resources, pages 33-40, 2014.

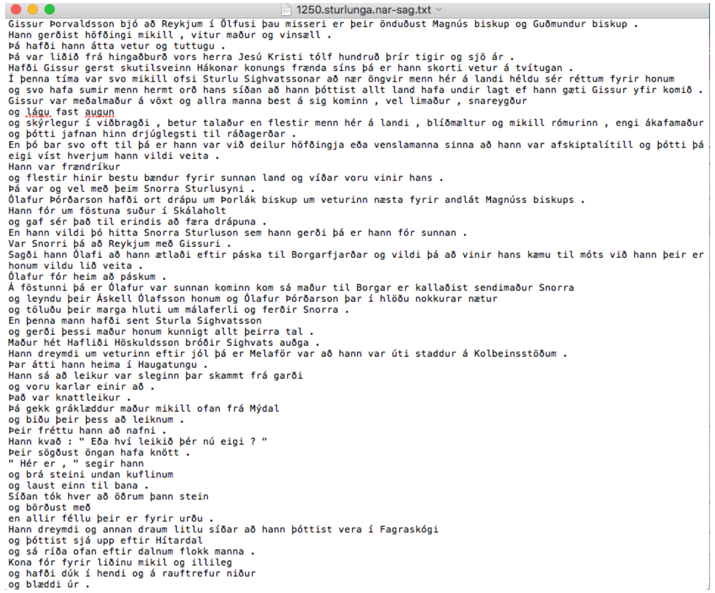
C. Schätzle, D. Sacha and M. Butt. Diachronic Visualization of Oblique Subjects in Icelandic. Workshop Poster: Big Data Visual Computing – Quantitative Perspectives for Visual Computing, 2014.

Diachronic Visualization of Text Properties/Icelandic

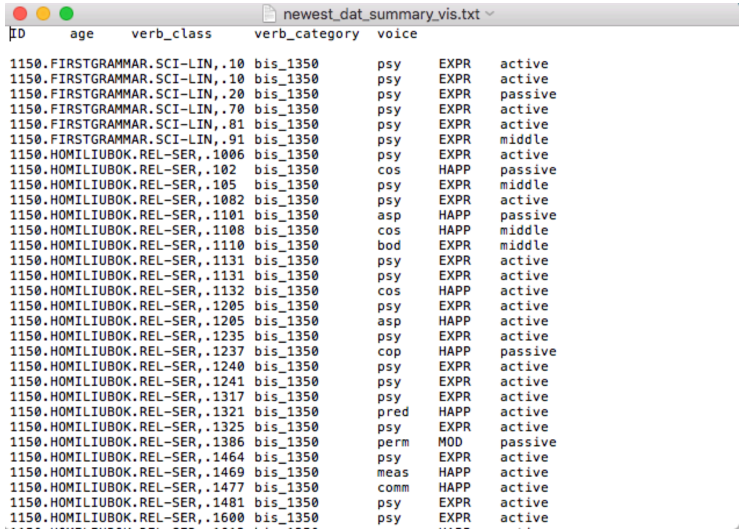
- You are provided with two tools for analyzing the the annotated diachronic corpus of Icelandic (IcePaHC).
- Two Questions so far:
  - When does V1 in Icelandic Occur?
  - What governs the appearance of dative subjects.
- Two Tools
  - icelandicV1 – Jar
  - Icelandic_dative-Jar
- Dataset is the same (data/text/)
- Derived Features/Properties (extracted sentences) are different

Check out the data

- Navigate into the data folder
- Take a look at the raw texts
- And the extracted language properties



```
Gissur Þorvaldsson bjó að Reykjum í Ölfusi þau misseri er þeir önduðust Magnús biskup og Guðmundur biskup .
Hann gerðist höfðingi mikill , vitur maður og vinsæll .
Þá hafði hann átta vetur og tuttugu .
Þá var liðið frá hingaðburð vors herra Jesú Kristi tólf hundruð þrír tigir og sjö ár .
Hafði Gissur gerst skutilsveinn Hákonar konungs frænda síns þá er hann skorti vetur á tvítugan .
Í þenna tíma var svo mikill ofsi Sturlu Sighvatssonar að nær öngvir menn hér á landi héldu sér réttum fyrir honum
og svo hafa sumir menn hermt orð hans síðan að hann þóttist allt land hafa undir lagt ef hann gæti Gissur yfir komið .
Gissur var meðalmaður á vöxt og allra manna best á sig kominn , vel limaður , snareygður
og lágu fast augun
og skýrlegur í viðbragði , betur talaður en flestir menn hér á landi , blíðmæltur og mikill rómurinn , engi ákafamaður
og þótti jafnan hinn drjúglegsti til ráðagerðar .
En þó bar svo oft til þá er hann var við deilur höfðingja eða venslamanna sinna að hann var afskiptalítill og þótti þá
eigi víst hverjum hann vildi veita .
Hann var frændríkur
og flestir hinir bestu bændur fyrir sunnan land og víðar voru vinir hans .
Þá var og vel með þeim Snorra Sturlusyni .
Ólafur Þórðarson hafði ort drápu um Þorlák biskup um veturinn næsta fyrir andlát Magnúss biskups .
Hann fór um föstuna suður í Skálaholt
og gaf sér það til erindis að færa drápuna .
En hann vildi þó hitta Snorra Sturluson sem hann gerði þá er hann fór sunnan .
Var Snorri þá að Reykjum með Gissuri .
Sagði hann Ólafi að hann ætlaði eftir páska til Borgarfjarðar og vildi þá að vinir hans kæmu til móts við hann þeir er
honum vildu lið veita .
Ólafur fór heim að páskum .
Á föstunni þá er Ólafur var sunnan kominn kom sá maður til Borgar er kallaðist sendimaður Snorra
og leyndu þeir Áskell Ólafsson honum og Ólafur Þórðarson þar í hlöðu nokkurar nætur
og töluðu þeir marga hluti um málaferli og ferðir Snorra .
En þenna mann hafði sent Sturla Sighvatsson
og gerði þessi maður honum kunnigt allt þeirra tal .
Maður hét Hafliði Höskuldsson bróðir Sighvats auðga .
Hann dreymdi um veturinn eftir jól þá er Melaför var að hann var úti staddur á Kolbeinsstöðum .
Þar átti hann heima í Haugatungu .
Hann sá að leikur var sleginn þar skammt frá garði
og voru karlar einir að .
Það var knattleikur .
Þá gekk gráklæddur maður mikill ofan frá Mýdal
og biðu þeir þess að leiknum .
Þeir fréttu hann að nafni .
Hann kvað : " Eða hví leikið þér nú eigi ? "
Þeir sögðust öngan hafa knött .
" Hér er , " segir hann
og brá steini undan kuflinum
og laust einn til bana .
Síðan tók hver að öðrum þann stein
og börðust með
en allir féllu þeir er fyrir urðu .
Hann dreymdi og annan draum litlu síðar að hann þóttist vera í Fagraskógi
og þóttist sjá upp eftir Hítardal
og sá ríða ofan eftir dalnum flokk manna .
Kona fór fyrir liðinu mikil og illileg
og hafði dúk í hendi og á rauftrefur niður
og blæddi úr .
```

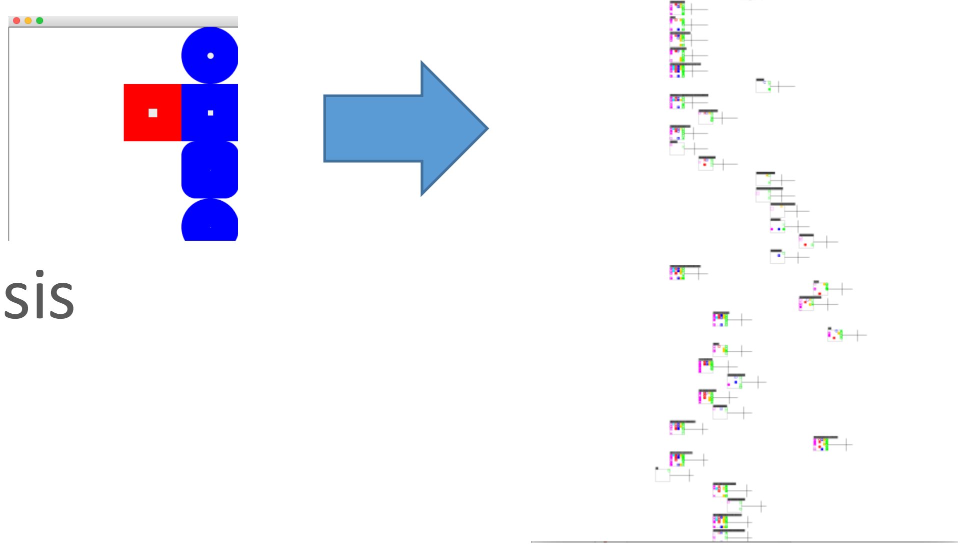


```
ID         age       verb_class        verb_category  voice

1150.FIRSTGRAMMAR.SCI-LIN,.10 bis_1350        psy    EXPR    active
1150.FIRSTGRAMMAR.SCI-LIN,.10 bis_1350        psy    EXPR    active
1150.FIRSTGRAMMAR.SCI-LIN,.20 bis_1350        psy    EXPR    passive
1150.FIRSTGRAMMAR.SCI-LIN,.70 bis_1350        psy    EXPR    active
1150.FIRSTGRAMMAR.SCI-LIN,.81 bis_1350        psy    EXPR    active
1150.FIRSTGRAMMAR.SCI-LIN,.91 bis_1350        psy    EXPR    middle
1150.HOMILIUBOK.REL-SER,.1006 bis_1350        psy    EXPR    active
1150.HOMILIUBOK.REL-SER,.102  bis_1350        cos    HAPP    passive
1150.HOMILIUBOK.REL-SER,.105  bis_1350        psy    EXPR    middle
1150.HOMILIUBOK.REL-SER,.1082 bis_1350        psy    EXPR    active
1150.HOMILIUBOK.REL-SER,.1101 bis_1350        asp    HAPP    passive
1150.HOMILIUBOK.REL-SER,.1108 bis_1350        cos    HAPP    middle
1150.HOMILIUBOK.REL-SER,.1110 bis_1350        bod    EXPR    middle
1150.HOMILIUBOK.REL-SER,.1131 bis_1350        psy    EXPR    active
1150.HOMILIUBOK.REL-SER,.1131 bis_1350        psy    EXPR    active
1150.HOMILIUBOK.REL-SER,.1132 bis_1350        cos    HAPP    active
1150.HOMILIUBOK.REL-SER,.1205 bis_1350        psy    EXPR    active
1150.HOMILIUBOK.REL-SER,.1205 bis_1350        asp    HAPP    active
1150.HOMILIUBOK.REL-SER,.1235 bis_1350        psy    EXPR    active
1150.HOMILIUBOK.REL-SER,.1237 bis_1350        cop    HAPP    passive
1150.HOMILIUBOK.REL-SER,.1240 bis_1350        psy    EXPR    active
1150.HOMILIUBOK.REL-SER,.1241 bis_1350        psy    EXPR    active
1150.HOMILIUBOK.REL-SER,.1317 bis_1350        psy    EXPR    active
1150.HOMILIUBOK.REL-SER,.1321 bis_1350        pred   HAPP    active
1150.HOMILIUBOK.REL-SER,.1325 bis_1350        psy    EXPR    active
1150.HOMILIUBOK.REL-SER,.1386 bis_1350        perm   MOD     passive
1150.HOMILIUBOK.REL-SER,.1464 bis_1350        psy    EXPR    active
1150.HOMILIUBOK.REL-SER,.1469 bis_1350        meas   HAPP    active
1150.HOMILIUBOK.REL-SER,.1477 bis_1350        comm   HAPP    active
1150.HOMILIUBOK.REL-SER,.1481 bis_1350        psy    EXPR    active
1150.HOMILIUBOK.REL-SER,.1600 bis_1350        psy    EXPR    active
```

Starting the Visualization

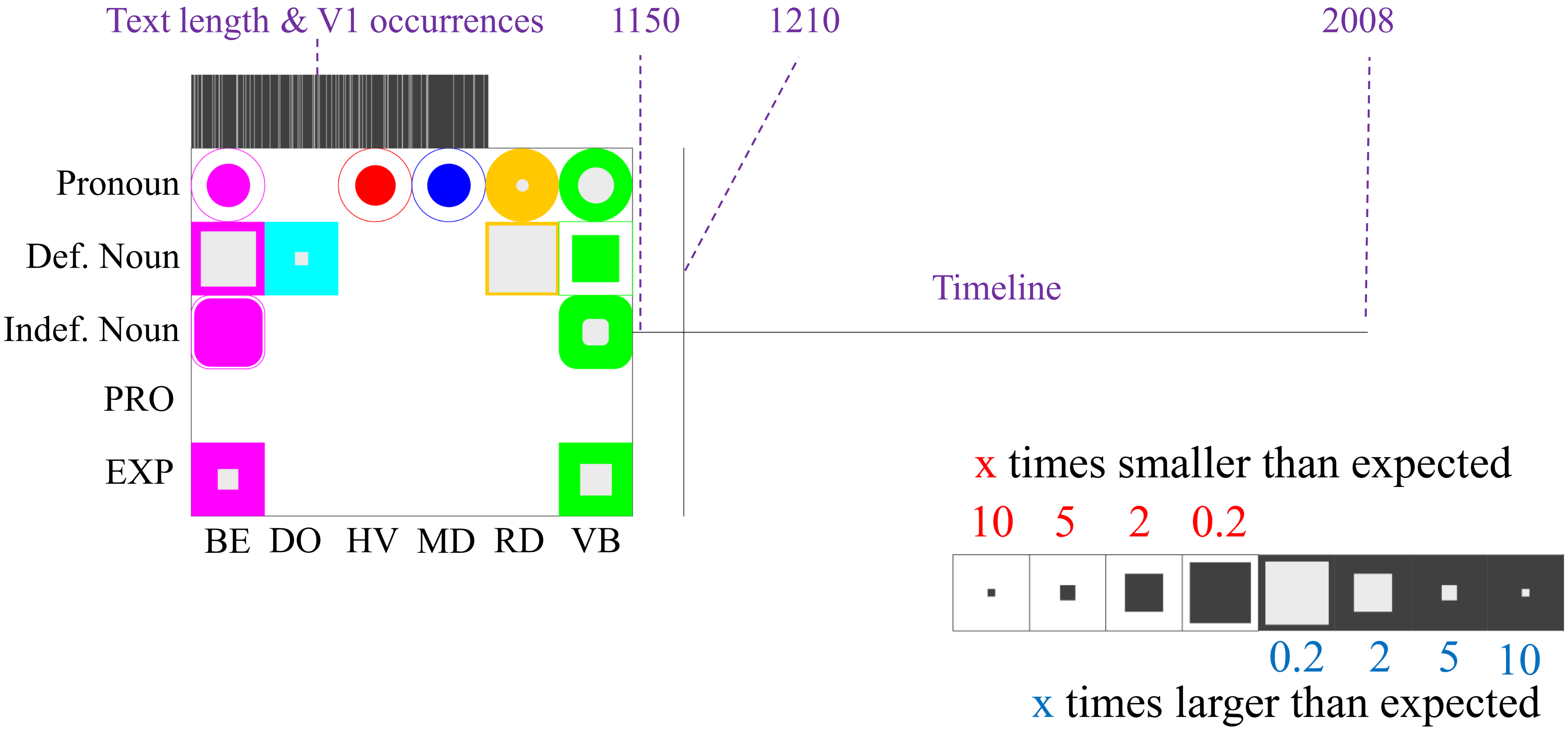
- To start the tools, just double click on the JAR-files
  - icelandic-1.0-SNAPSHOT-jar-with-dependencies.ja
  - dative-1.0-SNAPSHOT-jar-with-dependencies.jar
  - (Two separate folders)
- A Window should open



- Enlarge the window, and start your analysis
  - Zoom (mouse wheel)
  - Pan/Navigate
  - Details on Demand (tooltips)
  - Further interaction are described on top of the genres (for dative vis)

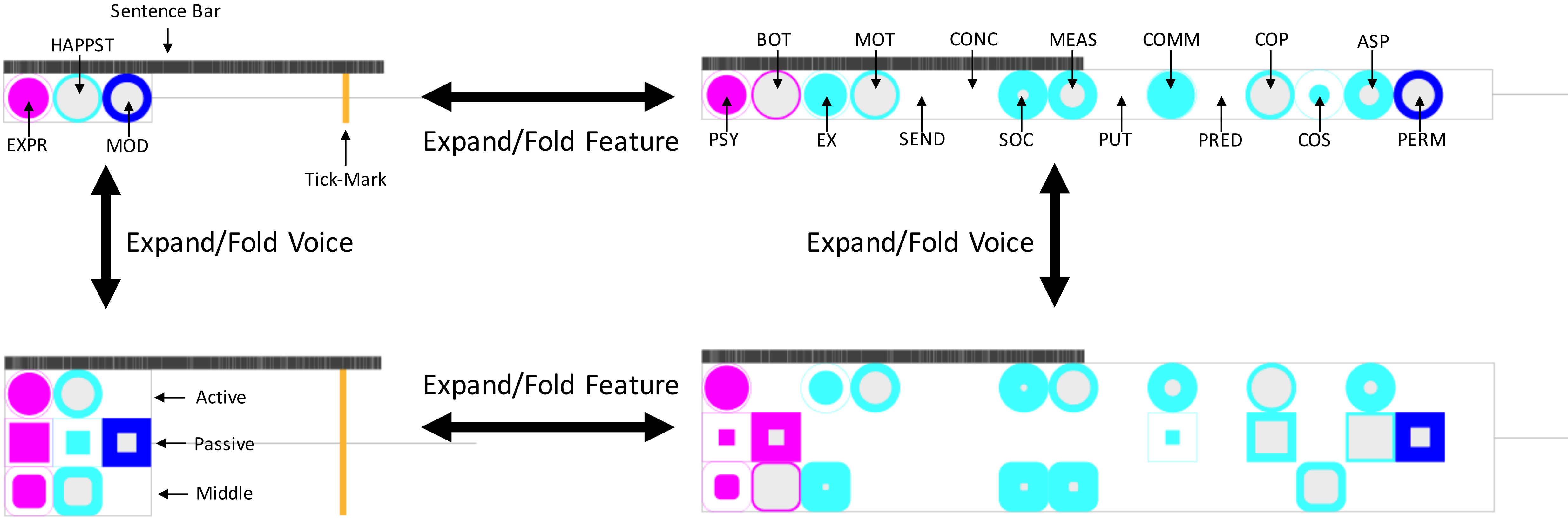
V1 in Icelandic

Factors identified by linguists as being relevant to V1 in Icelandic



Dative in Icelandic – Expand/Fold

- Press “e” or “t” to expand/fold the text nodes
- Clicking on the texts is also possible



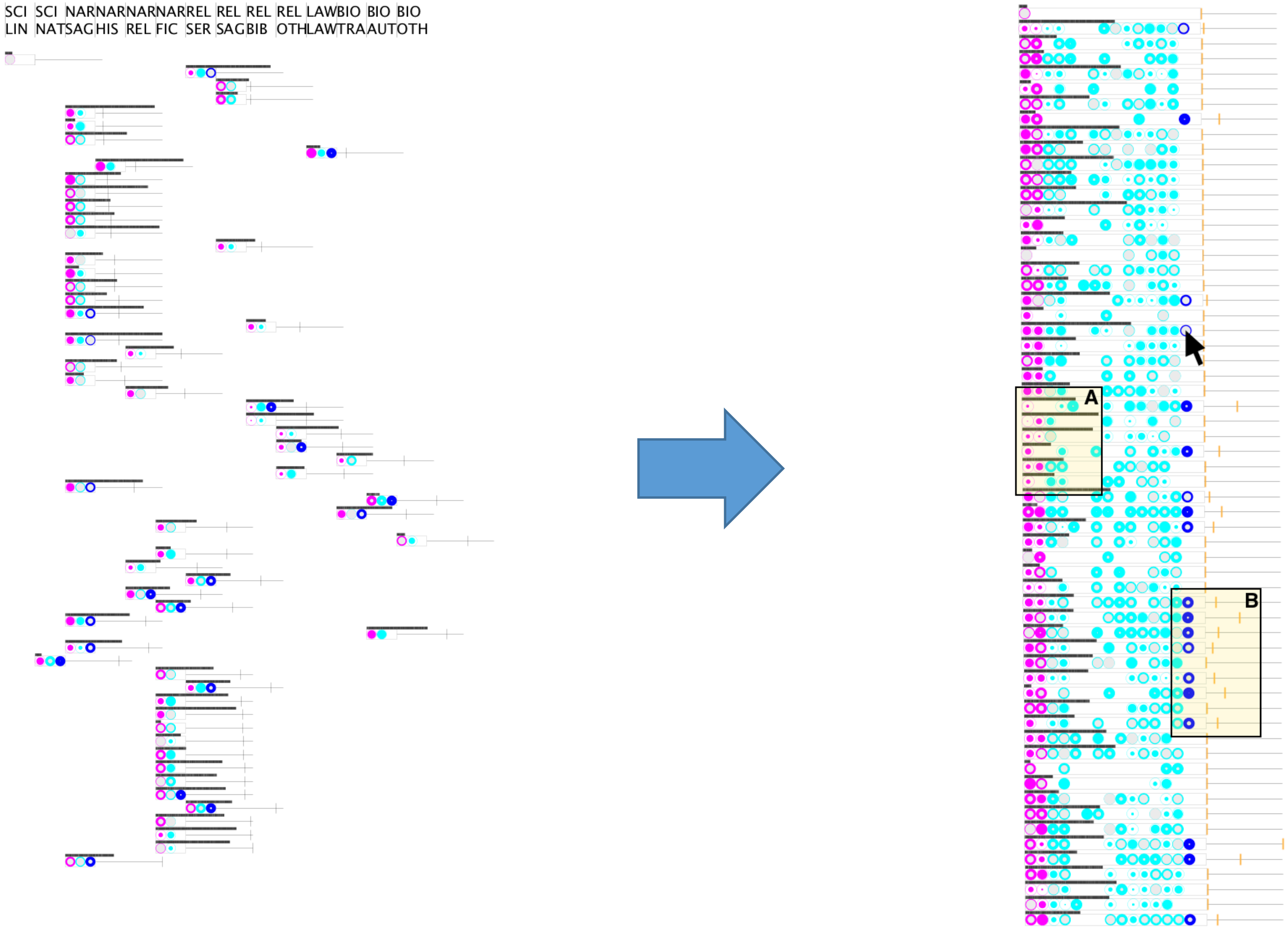
Expand/Fold Feature

Expand/Fold Voice

Expand/Fold Voice

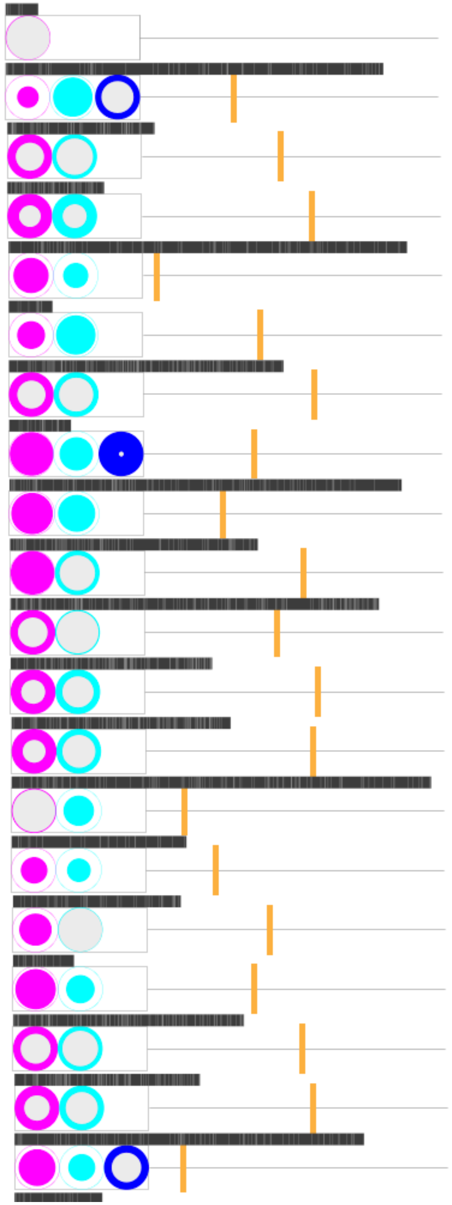
Expand/Fold Feature

Dative in Icelandic – Layout (Press “l”)



Dative in Icelandic – Interactive Tick-Marks

- Press “d” to disable the tooltips
- Press “l” to change the layout
- Hover over a glyph to activate its tick marks



Hints – It is a Research Prototype!

- If the program hangs (only white space is shown)
  - →Just restart the app (open a new window)

- Task Suggestion
  - Work with the software as is.
  - Think critically about the visualization/interactive possibilities.
  - See if you can identify patterns from the visualization without necessarily knowing anything about Icelandic or the phenomenon (we could).

Cluster Visualization

Various Perspectives on Data

Andreas Lamprecht, Annette Hautli, Christian Rohrdantz, Tina Bögel. 2013. A Visual Analytics System for Cluster Exploration. *In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, System Demo,* 109–114, Sofia, Bulgaria.

Cluster Visualization

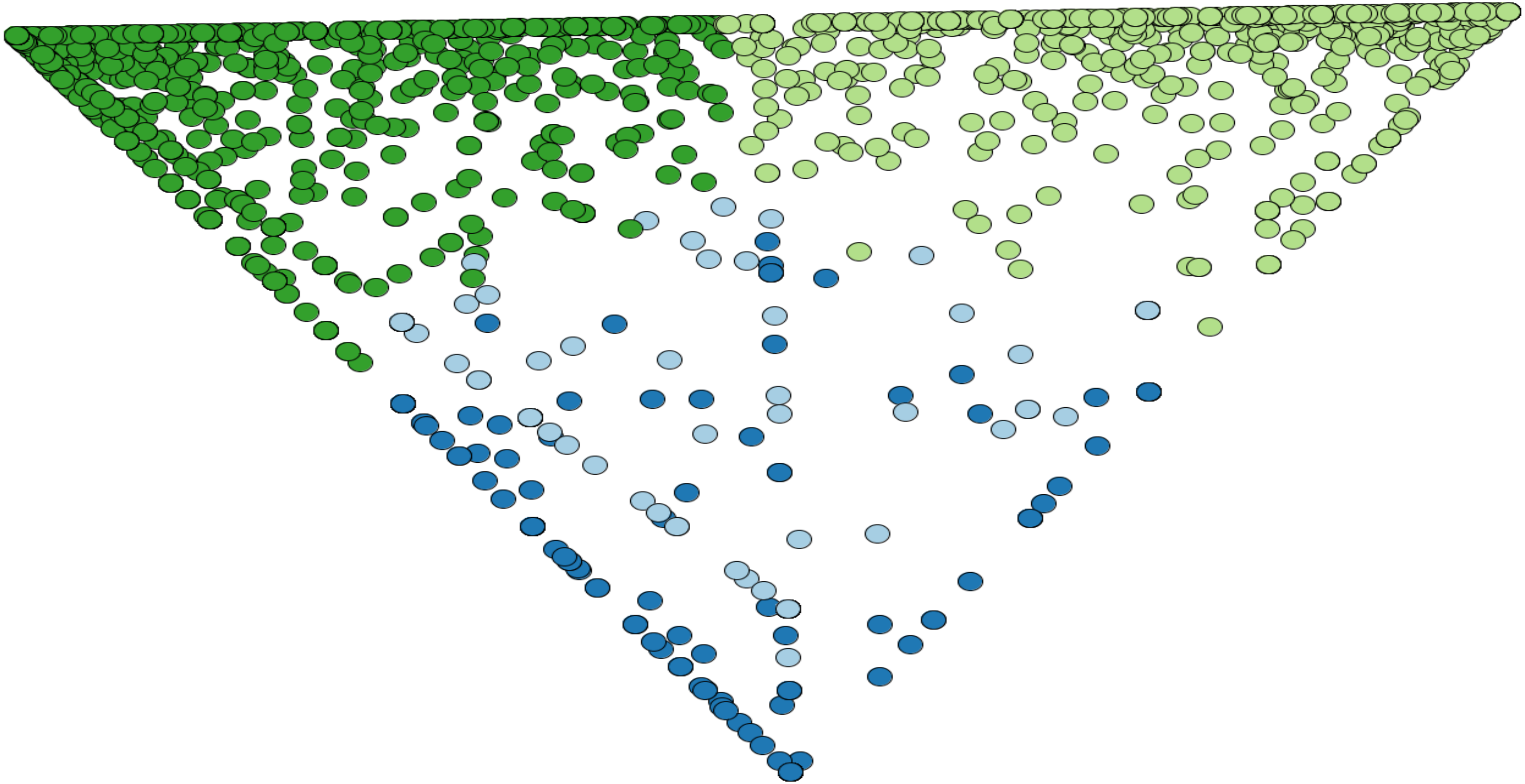
- Automatic clustering methods are increasingly being used by a wide range of linguists.
- However, it is often hard to understand what the clustering method is doing
- And it is hard to interact with it.
- The following presents an interactive, flexible visual analytic approach to clustering information.

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

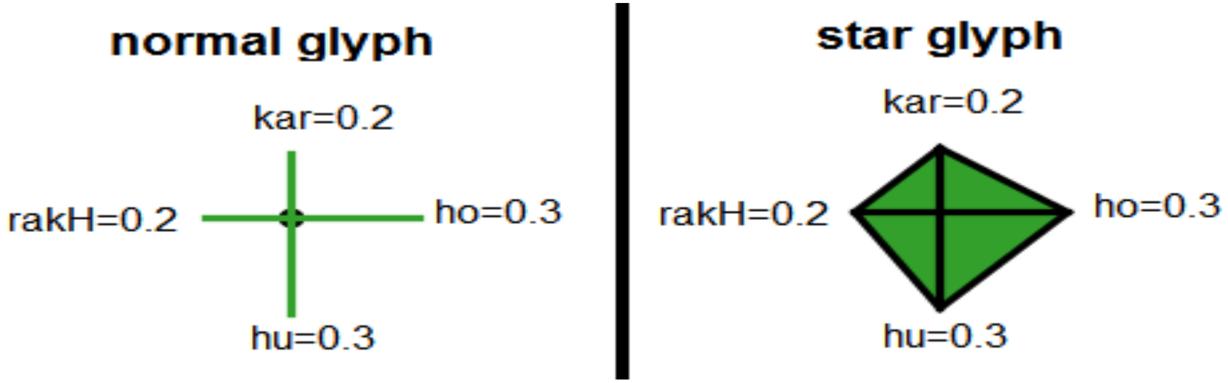


Glyphs in the Cluster Visualization

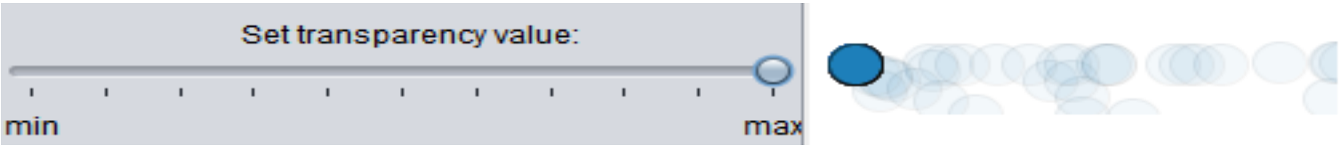
- Glyphs are combinations of symbols that are defined to have a certain meaning.
- Data objects in the visualization can be presented either as circles, normal glyphs or star glyphs.
    - Circles: Every items (e.g. a noun or a verb) represented by a colored circle
    - Normal glyphs: Relative bigram 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".

Glyphs in the Cluster Visualization

- The data shown here is that from the N-V complex predicates case study.
- 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.

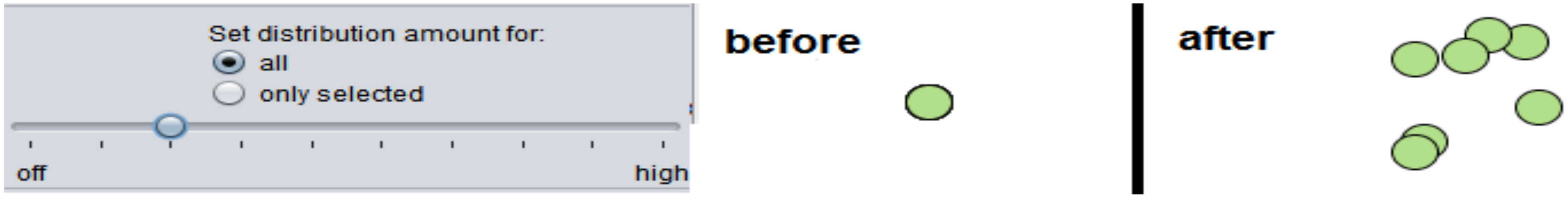


- Overplotting is a problem when the data set becomes large or when the data points are very similar to one another.
- Several strategies to handle this interactively:
  - change transparency of objects



Set transparency value:

  - reposition data objects

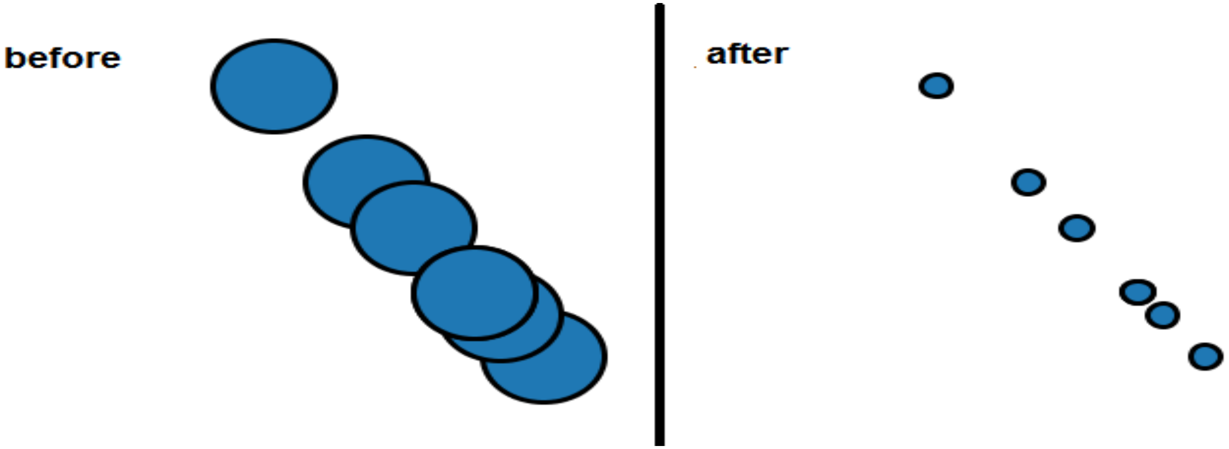


Set distribution amount for:

● all

○ only selected

  - scale data objects



Cluster Visualization

- You are provided with a Java program in class.
- The software is still under development, so if you want to use it for purposes outside of this class, please contact Miriam.
- It should start by just clicking on it.
- There is a **Readme** file to guide you through what needs to be done.
- There is also an extensive handbook in PDF format.

Cluster Visualization

Data:

- The data needs to be in a txt file.
- The data points need to be separated by a symbol (e.g. “,”)
- We have provided sample data from our work on Urdu
  - Motion verbs courtesy of Annette Hautli – this seems to be broken on my version.
  - Urdu N-V complex predicates
- We have also provided some data based on Levin’s verb classes (levin-classes.txt).
- Feel free to add to this data as you wish.
- Some information on Levin’s verb classes is provided in levin-verbs-lawler.txt.

Cluster Visualization

Task/Interaction Suggestion:

- work with the Urdu motion verbs or the Levin verb classes file to get a feel for the visualization
- experiment with different numbers of clusters
- experiment with different visualizations of the data points (glyphs, star glyphs)
  - the Levin verb classes file contains three errors (three verbs contain wrong information)
  - see if you can spot that via the visualization
- think critically about the visualization and the interactive possibilities

Cluster Visualization

Task/Interaction Suggestion:

- enter your own data into a file by using the existing ones as a model
- you need to think about how to encode your data so that the system can compute with it
- Example: you may be interested in properties like whether a noun takes a certain case marker
  - Noun1: accusative, instrumental
  - Noun2: accusative, no instrumental
- this can be encoded as:

```
NounType, accusative, instrumental
Noun1, 1, 1
Noun2, 1, 0
```

ClusterVis without verbs

- **Idea:** Sometimes we can use visualizations with data other than what they were originally designed for.

- **Example:** ClusterVis was designed to analyse properties of verbs, but it can be used to analyze any similarly encoded properties, no matter what those properties are for.

- **To try:** bierce-freq.txt and bierce-freq-2.txt contain information about letters that Ambrose Bierce wrote. The features include things like the number of pronouns (PP), and the number of words longer than 6 characters. Are there any clusters? If so, can you interpret them? (The data originally are from Chris Culy and we don't know the answers....)

The WALS Sunburst Explorer

Visualization of Typological Patterns

Christian Rohrdantz, Michael Hund, Thomas Mayer, Bernhard Wälchli and Daniel A. Keim. 2012. The World’s Languages Explorer: Visual Analysis of Language Features in Geneaologica and Areal contexts. *Computer Graphics Forum* 31(3), 935-944.

WALS Explorer

- The WALS Explorer can be accessed on-line:
  - http://th-mayer.de/wals/
- You cannot use your own data here.
- It is meant for an exploration of the World Atlas of Language Structure (http://wals.info).
- Task/Interaction Suggestion:
  - pick a phenomenon you are interested in
  - see what you can find out about it
  - think critically about the visualization and the interactive possibilities

DoubleTree

Visualizations by Chris Culy

Exploring Texts

Chris Culy has designed a number of visualization possibilities

- http://linguistics.chrisculy.net/lx/software/
- You can either download these or use them on-line

Suggestions:

Explore the examples provided with the visualizations

- What are advantages/disadvantages of each?
- What would you like them to do that they can't?

- If you have tagged data available, try your own.

Some Further Available Visualization Tools/Ideas

- http://magic-table.googlecode.com/svn/trunk/magic-table/google_visualisation/example_1.html

- http://www.eurac.edu/en/research/autonomies/commul/projects/Pages/Linfovis/programs.aspx

- http://www.knime.org/

Exploring bigrams with MagicTable
Medium-advanced: basic programming

http://magic-table.googlecode.com/svn/trunk/magic-table/google_visualisation/example_1.html

- Use the MagicTable visualization from Google charts to look at bigram co-occurrences
    - cell row,column is for the bigram: row column

- Data:
    - maybe look at POS tag bigrams
    - have to count and normalize