A Visual Analytics System for Cluster Exploration

Andreas Lamprecht[1], Annette Hautli[2], Christian Rohrdantz[1], Tina Bögel[2]

[1]Department of Computer Science, [2]Department of Linguistics, Universität Konstanz, Germany

Motivation

- Visual analytics systems are increasingly used for the investigation of linguistic phenomena.
- Interpretability of results coming from machine learning algorithms is an issue in computational linguistics.
- Insights into cluster constituency and prototypical cluster members (centroids).

Aims

- Present a visual analytics system which facilitates "analytical reasoning by an interactive visual interface".
- Present strategies to deal with a large number of data points.
- Get an at-a-glance overview of the statistical exploration of a linguistically motivated phenomenon.

Case study

- Linguistic phenomenon: The varied behavior of nouns in Urdu N+V complex predicates **??**.
- Complex predicates in English: *take a nap*, *get rid of*
- In Urdu: CPs as a preferred way of expressing events.
- Urdu nouns can take a range of *light verbs* that change the interpretation of the CP:

(1) yad kɪ-ya
memory do-Perf.M.Sg
'to (actively) remember sth.'

(2) yad hɛ
memory be-Perf.Sg
'to (passively) remember sth.'

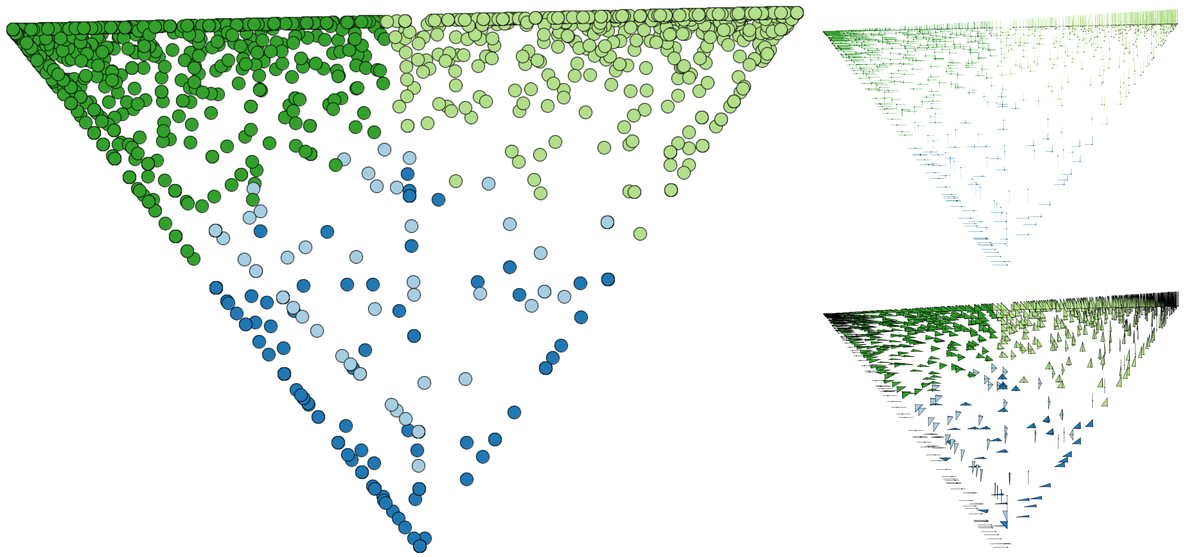
(3) yad hu-a
memory be.Part-Perf.M.Sg
'to come to remember sth.'

- Investigated light verbs: *kar* 'do', *ho* 'be', *hu* 'become' and *rakʰ* 'put')
- Not all nouns can take all light verbs in a CP → Hints towards a semantic difference between the nouns.

The visual analytics system

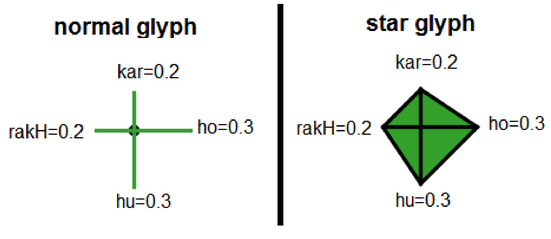
- Input: bigram frequencies (absolute or relative) of noun + light verbs extracted from the Urdu BBC corpus.
- Initial clustering calculated in the high dimensional space using a k-Means algorithm.
- Projection onto the two-dimensional space using a Principal Component Analysis (PCA) algorithm.
- Each data point represents one noun and its light verb behavior.

Different visualizations of data points



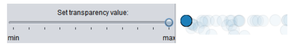
Visualization types

- Data objects are 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"



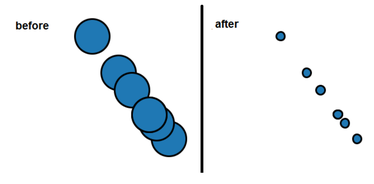
Handle overplotting

- Change the transparency of data objects:



- Reposition data objects:

- Scale data objects:



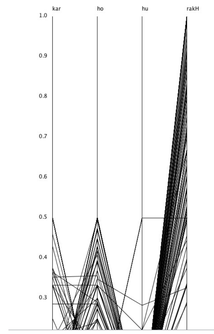
The interactivity of the system

- **Filtering:**
  - Bigram frequency: E.g. only show nouns which occur with selected features (light verbs) exclusively
  - Frequency above a certain threshold: E.g. show nouns which exceed a defined minimal frequency in the considered corpus.
  - Filter by cluster/class: show only a selected cluster/class
- A specific group of data points can be **selected, inspected, extracted, re-clustered, re-visualized and stored** using the visualization system.
- The system allows to **zoom in and out** of the cluster visualization → find patterns based on different perspectives on the data.

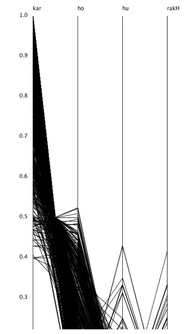
Alternative views

- Correlation matrix: shows the correlation between every pair of features (here: light verbs).
- Scatter plot matrix: shows a scatter plot for every feature (here: light verb) combination.
- Parallel coordinates: shows the distribution of the bigram frequencies across the different features.

Cluster 1:



Cluster 4:



Benefits of the visualization

- Facilitation of hypothesis-testing and -generation by representing data visually.
- Insights into the "black box" of clustering: constituency of the cluster, prototypical cluster members, distance of each data point to the centroid.
- System provides interpretable results which eases the communication between researchers of different fields.
- Built-in options and add-ons are designed for the type of work linguists are interested in: overview first, in-depth data inspection later.
- The interactivity allows for new interpretations of the data.

References

Ahmed, Tafseer and Butt, Miriam. 2011. Discovering Semantic Classes for Urdu N-V Complex Predicates. In *Proceedings of the international Conference on Computational Semantics (IWCS 2011)*, pages 305–309.

Butt, Miriam, Bögel, Tina, Hautli, Annette, Sulger, Sebastian and Ahmed, Tafseer. 2012. Identifying Urdu Complex Predication via Bigram Extraction. In *In Proceedings of COLING 2012, Technical Papers*, pages 409 – 424, Mumbai, India.

Acknowledgments

This work was partially funded by the German Research Foundation (DFG) under grant BU 1806/7-1 "Visual Analysis of Language Change and Use Patterns" and the German Federal Ministry of Education and Research (BMBF) under grant 01461246 "VisArgue".