



LingVis: Motivation and Use Cases

Miriam Butt & Dominik Sacha

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



Federal Ministry of Education and Research

Tutorial Overview

- Part I: Motivation and Use Cases
- Part II: Background Theory
- Part III: Hands-on
 - Presentation of Possibilities
 - More background on some of the systems
 - Chance to work/explore yourselves

Before we start....

Personal Questions

- Who are we?
- Who are you?
 - Programming Background
 - What types of linguistic questions interest you?
 - What is the interest in LingVis?

Motivation and Background

LingVis

Overall Goals:

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



Visual Analytics

- Interactive, exploratory access to data
- Iterations of hypothesis formation and hypothesis testing
- Overview first details on demand
 - Holistic picture
 - But can drill down to individual data points
 - Abstract representation of multifactorial, multidimensional data.
 - Good for understanding complex interactions in the data.

Sample Visualizations



- Linguists are making more and more use of newly available technology to detect **distributional patterns** in language data.
- Ever increasing availability of **digital corpora** (synchronic and diachronic).
- Increasing interest in language output produced in social media.
- Ever better query and search tools (CQP, COSMAS, DWDS, ANNIS).
- **Programming languages** suitable for text processing, statistical analysis and visualization (e.g., Python, R).
- But: as yet only comparatively little/good use of novel visualization methods.

Making Sense of Numbers

- Current linguistics often includes corpus work.
- Linguists try to determine patterns, interactions and usage preferences within a language but also across different languages.
- This work generates a lot of numbers (statistics).
- Numbers are difficult for humans to process.
- Solution: translate numbers into visual properties.
- Human visual apparatus can process this easily.

Interdisciplinary Collaboration: LingVis

Research Question

Data / Language Resources

Domain Expert

Interdisciplinary Collaboration: LingVis

Research Question



Interdisciplinary Collaboration: LingVis

Research Question



Example: Pixel-Based Visualizations

Two Use Cases

- N-V Complex Predicates
- Vowel Harmony

N-V Complex Predicates

- N-V complex predicates occur very frequently in Urdu.
- Examples: phone-do, memory-do, memory-become, resolution-do, resolution-be, ...
- **Problem:** would be nice if one knew which nouns were likely to cooccur with which verbs.

Example: N-V Complex Predicates in Urdu

- **Goal:** identify sequences of Noun+Verb for understanding complex predicate patterns
 - phone-do, use-do, memory-come, begin-do/come
- Data: 7.9 million word raw (unannotated) corpus of Urdu (BBC Urdu)

1	#this file	lists	X in X	(+kar, X-	ho, X+hu, X+rakh sequences with co	or
	espondin	g occ	urren	ces in th	e (candidate) CP sequences	
2	#X = wor	d occ	urring	g directly	to the left of LV (LV: kar, ho, hu, rak	h)
3	#kar: # 0	foccu	urrend	ces of X v	vith kar	
4	#ho: # of	occu	rrenc	es of X w	ith ho	
5	#hu: # of	occu	rrenc	es of X w	ath hu	
6	#rakh: #	of oco	currer	ices of X	with rakh	
	A	#nu	#Kar	#no #ra		
8	مفاص	674	466	524	0	
9	عودش	378	2330	5 1691	0	
10	مولعم	366	254	609	0	
11	_م ڪامھد	359	135	44	0	
12	Cal	227	1232	2 100	0	
13	رثاتم	183	178	765	0	
14	ناصقن	173	0	114	0	
15	ایک	172	373	7027	0	
16	تباث	147	394	588	0	
17	تقو	142	105	235	9	
18	ادىپ	103	754	956	0	
19	کالہ	102	150	1 3609	0	
20	دمآرب	80	210	96	0	
21	اهکر	74	0	263	0	
22	ىمئز	62	59	1161	0	
23	زأعآ	59	315	75	0	
24	~	56	0	2267	0	
25	دقعنم	54	197	262	0	
26	فاشكنا	51	165	13	0	

Butt, Miriam, Tina Bögel, Annette Hautli, Sebastian Sulger & 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.

Example: Pixel Visualization

Statistical Data:

ID	Noun	Rel. freq. with kar	Rel. freq. with ho	Rel. freq. with hu	Rel. freq. with $r \alpha k^h$				
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





Pixel plus Cluster Visualization

- Performed k-means clustering combined with a pixel visualization.
- Advantages:
 - can inspect clusters visually and detect patterns
 - Outliers spotted easily (mostly errors "kyA" is not a noun, it is a *wh*-word and was included by mistake).





Example: Identifying N-V complex predicates in Hindi/IUrdu

Tool facilitates zooming and mousing over to see the underlying data set



Outliers/Errors are easily identified (Clustering Algorithm has applied)



N-V Complex Predicates

Cluster Visualization Demo

More sophisticated version now available – will look at that in hands-on part

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.

Improved Cluster Visualization (Lamprecht et al.)



Vowel Harmony (VH)

- Phenomenon (simplified): Vowels in affixes change according to vowels found in stems.
- (Famous) Example: Turkish

Genitive suffix

deniz-in, ev-in tütün-**ün**, çöl-**ün** kadın-ın, adam-ın sabun-**un**, top-**un**

Genitive suffix with plural suffix deniz-ler-in, ev-ler-in tütün-ler-in, çöl-ler-in kadın-lar-ın, adam-lar-ın sabun-lar-ın, top-lar-ın

Vowel Harmony

Goal: Try to determine automatically whether a given language contains patterns indicative of vowel harmony.

Basic Computational Approach:

- Use written corpus (caveat: only approximates actual phonology).
- Count which vowels succeed which other vowels in VC⁺V sequences (within words — again an approximation)
- Through **statistical analysis** find out the association strength between vowels: normalized association strength value ϕ .
- **Results** show that Turkish and Hungarian, for example, pattern similarly. Languages like Spanish or German pattern differently.

Results — Standard Methods: Can you detect a pattern?

	а	I	u	0	ö	ü	i	е		a	i	0	e	u				
а	0.266	0.427	-0.141	-0.060	0.019	-0.125	-0.261	-0.275	a	-0.003	-0.075	0.094	-0.025	-0.018				
I	0.162	0.292	-0.107	0.077	-0.010	-0.075	-0.190	-0.191	i	-0.025	-0.004	0.064	-0.036	0.005				
u	0.129	-0.143	0.464	0.017	-0.003	-0.051	-0.138	-0.140	0	-0.028	-0.006	-0.075	0.098	0.026				
0	0.066	-0.112	0.434	-0.015	0.006	-0.045	-0.104	-0.111	e	-0.001	0.063	-0.073	0.016	0.021				
ö	-0.107	-0.092	-0.052	-0.026	0.006	0.366	-0.091	0.164		0,077	0.038	0.036	0.057	0.043				
ü	-0.120	-0.114	-0.059	0.014	-0.006	0.507	-0.112	0.134	- u = 0.077 = 0.038 = -0.030 = -0.037 = -0.04									
i	-0.201	-0.224	-0.118	0.071	-0.004	-0.087	0.319	0.211	Spanish									
е	-0.256	-0.251	-0.132	-0.062	-0.010	-0.097	0.400	0.276										

Turkish

				-						^	^				^	•	
	a	0	i	ü	Ö	ä	u	e		a	0	u		u	0	e	
a	0.019	0.009	-0.061	-0.034	-0.008	-0.025	0.018	0.035	a	0.339	0.263	0.070	-0.022	-0.081	-0.136	-0.431	
0	-0.023	-0.004	-0.052	-0.013	-0.020	-0.013	-0.013	0.068	0	0.239	0.099	0.041	-0.007	-0.052	-0.083	-0.253	
i	-0.069	-0.054	-0.050	-0.039	-0.036	-0.044	-0.003	0.133	u	0.132	0.038	0.015	-0.004	-0.017	-0.040	-0.131	
ü	-0.067	-0.045	0.070	-0.028	-0.021	-0.033	-0.021	0.050	i	0.037	-0.026	0.008	-0.030	-0.017	-0.027	0.011	
ö	-0.049	-0.032	0.049	-0.024	-0.013	-0.021	-0.013	0.036	ü	-0.093	-0.056	-0.022	-0.014	0.008	0 148	0.071	
ä	-0.067	-0.037	0.124	-0.033	-0.018	-0.028	-0.038	0.020		0.000	0.000	0.022	0.011	0.000	0.110	0.007	
u	0.012	-0.018	-0.019	0.046	-0.002	-0.013	0.004	-0.001	0	-0.152	-0.093	-0.037	0.001	0.065	0.229	0.097	
е	0.108	0.084	0.026	0.069	0.063	0.096	0.021	-0.195	е	-0.435	-0.241	-0.076	0.048	0.091	0.054	0.531	
		I	1	1	L	1											

German

Hungarian

...

First Simplistic Visualization: Can you detect a pattern?



- Matrix visualization of association strengths between vowels (deviation from statistical expectation).
- Vowels are sorted alphabetically.
- More saturated colors show greater association strength.
- Blue is for more frequently than expected, red for less.
- The +/- are redundant encodings.

Sorted Visualization: Can you detect a pattern now?



Vowels **sorted** according to similarity (note: not a trivial process) Can even see the **type** of Vowel Harmony involved.

T. Mayer, C. Rohrdantz, M. Butt, F. Plank and D. A. Keim. Visualizing Vowel Harmony. *Linguistic Issues in Language Technology*, 4(Issue 2):1-33, 2010.

Visualizing Vowel Harmony

Counting Vowel Successions	in all	Bible T	ypes
Example: Finnish			

	a	ä	e	i	0	ö	u	У	
a	3548	20	1940	1893	831	0	944	24	
ä	35	944	806	820	10	138	33	266	
e	1623	1144	1495	1608	419	56	497	187	
i	1580	854	1514	1044	376	46	355	135	
0	1384	7	1032	902	284	0	294	8	
ö	7	125	54	39	0	3	1	18	
u	1464	6	1085	850	315	1	547	8	
У	39	656	368	368	$\overline{35}$	$\overline{75}$	4	251	

Sorting done according to feature vectors of each of the rows.

Statistics & Visualization





Results – Sorted Visualization:

• Automatic Visual Analysis of vowel successions for 42 languages – sorted for effect strength.



Vowel Harmony vs. Reduplication

- In VH languages, crucially there are some vowels which never cooccur.
- This can be seen via a calculation of succession probabilities.
- Maori is not a VH language.



Historical Fingerprint: German Umlaut

Even though Umlaut (raising of vowel in stem before high vowel in affix) is no longer a productive process in German, the Umlaut harmony pattern is still visible in the matrices.



You can use the visualization in a new and improved form yourself on-line.

http://paralleltext.info/phonmatrix/

Main Contact Person: Thomas Mayer

Mayer, Thomas and Christian Rohrdantz. 2013. PhonMatrix: Visualizing co-occurrence constraints in sounds. In *Proceedings of the ACL 2013 System Demonstration*.

Example: Droplet Visualizations

- Different Types of Visualizations can be used to look at the same data.
- Example: Droplets for Vowel Harmony
- This droplet technique was originally used for rendering geospatial information (an item moving from one place to the next).

Vowel Harmony via Droplets

kaşık-lar-ım-a spoon-Pl-1SgPoss-Dat 'my spoons'

kedi-ler-im-e cat-Pl-1SgPoss-Dat 'my cat'



Language Comparison via Droplets



Norwegian shows language change a \rightarrow e in comparison to Swedish.

Example: Sunburst and Maps

- Another way to compare features across languages is via a sunburst visualization.
- The following visualization combines sunburst with a link to the geographical location of the language.
- The visual analysis is heavily interactive.
 - One can feed in one's own data.
 - One can also use the WALS (World Atlas of Language Structures; http://wals.info).

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.

Sunburst and Maps for Language Families



World's Language Explorer

Comparing 126 Languages of Papua New-Guinea based on the New Testament.



Each circle segment represents one language, each ring the values of one feature across all languages.

World's Language Explorer



Bringing genealogy (left) and areal distributions (right) interactively into context: The values of a selected feature ring are color-coded on a map for exploration.

Sorting and Pattern Discovery



Figure 2.5: High-resolution screenshot showing automatically extracted features for languages from Papua New Guinea with leaves ordered to *maximize* (left) and *minimize* (right) the pairwise leaf similarity for neighbors.

Sorting and Pattern Discovery



WALS Explorer

- A version that is tailored to interact with WALS is available on-line
- http://www.th-mayer.de/wals/

Thomas Mayer, Bernhard Wälchli, Christian Rohrdantz and Michael Hund. 2014. From the extraction of continuous features in parallel texts to visual analytics of heterogeneous areal-typological datasets. In B. Nolan and C. Periñán-Pascual (eds.), Language Processing and Grammars: The role of functionally oriented computational models, 13–38. John Benjamins.

Conceptual Recurrence Plots

- Another type of much studied language data: discourses.
- The context of social media (Twitter, Facebook, etc.) presents us with new opportunities but also with new challenges.
- Next up: visual analysis of a (conventional) dialog an interview.

Daniel Angus, Andrew E. Smith, Janet Wiles: **Conceptual Recurrence Plots: Revealing Patterns in Human Discourse**. *IEEE Trans. Vis. Comput. Graph.* 18(6): 988-997 (2012)







Fig. 9: Engagement Block (Denton/Kennett)

Fig. 10: Random Scattering (Denton/Singer)



Fig. 11: Concept Drift (Denton/Singer)

Glyph Visualization for Diachronic Corpora

- Visualization of IcePaHC Diachronic Corpus of Icelandic
 - Syntactically annotated in Penn Treebank style
 - 60 texts
 - 12th century CE to 21st century CE
- Two case studies so far (on-going)
 - V1 in Icelandic
 - Dative Subjects in Icelandic

Glyph Visualization for Diachronic Corpora

V1 (Verb Initial or Verb First)

- Verb initial structures were common in matrix declaratives in Germanic.
- In German (and English) they mostly survive in narrative/joke contexts
 Walked a man into a pub...

Questions

- What determines the appearance of V1?
- How did this change over the history of Germanic?

Butt, Miriam, Tina Bögel, Kristina Kotcheva, Christin Schätzle, Christian Rohrdantz, Dominik Sacha, Nicole Dehé & 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). Reykjavi Iceland.

Example: V1 in Icelandic

Visual Analytic Access to Data

- Glyph Visualization of likely factors
- Overview of all 60 texts at once
- Can drill down to individual data points interactively
- Keim's Mantra: Overview First, Show the Important Details on Demand





Icelandic Visualization Demo

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.

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)

Dominik Sacha, Yuki Asano, Christian Rohrdantz, Felix Hamborg, Daniel A. Keim, Bettina Braun & Miriam Butt. 2015. Self Organizing Maps for the Visual Analysis of Pitch Contours. Proceedings of the 20th Nordic Conference of Computational Linguistics (NoDaLiDa-2015), Vilnius, Lithuana, 2015.

Example: Speech Data via SOMs

- Japanese Native and German L2 Learner data (pitch contours and meta data)
- F0 contours are smoothed and normalized into pitch vectors
- The pitch vectors are visualized via self-organizing maps (SOM)

Speakers pronounced "*sorry/excuse me*" in ever more exasperating circumstances

- Japanese natives do not vary the pitch contour
- German learners do vary the pitch
- German beginner learners do so more

Interactive Exploration:

- individual cells can be merged
- meta data can be inspected
- individual pitch tracks can be examined in context



German Entschuldigung 'sorry' vs. Japanese Sumimasen 'sorry'



Example: Speech Data via SOMs

• Different views on the data and meta data can be explored interactively



Figure 3: Different approaches to visualize SOM-results according to available meta-data. (A) Grid visualization, (B) word cloud, (C) bar charts, (D) mixed color cells, (E) ranked group clusters, (F) one single cell that visualizes contained vectors and the cluster prototype, (G) separated heatmaps for all values of a categorical attribute.

Interactive Data Exploration: SOMs in Action



Configuration Cell Interactions

Filtering

(Re-)Training



Data Exploration

- SOM learning is fast
- user can switch among different perspectives on the data
- user can interactively delete or pin cells
- and retrain and re-explore

Identifying Optimal Visualizations

- Understanding which visualizations are optimal is not trivial.
- Are the individual dimensions (color, shape, direction, size, etc.) usefully meaningful?
- Does the visualization allow for at-a-glance understanding, or does it confuse the user?
- This also depends on the user's background
 - What is the user used to looking at?
 - How is the user used to understanding the data?
 - How is the user used to interacting with systems?
- Currently evaluations are mainly performed via user studies in Visual Analytics.
- New Project (SFB/TRR 161): Establish evaluation metrics.

Distorted Map according to number of languages spoken in area.



Visualization only as good as your data – India massively underrepresented

Using Motion

 highly frequent words in New York Times articles 2004-2005 and their relation to one another

tsunami indonesia

show trends/change

earthquake Challenges for Visualization: dimensionality reduction: war disaster high dimensional distance matrices iraq bush are shown in 2D louisiana precision vs. stability: florida a precise visualization for each hurricane time step would induce too much confusing movement 2004-11

(Work Group: Oliver Deussen, University of Konstanz)

Example: Animated Visualization

- the raw data without visualization:
 - 9x9 distance matrices for each of the 14 time steps

			×		disaste	r	indone	sia	earthqu	ake	bush	war		iraq	1	ouisiana	hur	ricane		florida	ı
			disas	ster		0	0.9	16525	0.9	11134	0.81982	3 0.	861211	0.869	453	0.97091	2	0.724	439	0.843196	1
			indor	nesia	0.9	16525	5	0)	1	0.96266	3 0.	945219	0.939	384	0.93972	3		1	0.978338	1
			earth	nquake	0.9	11134	F	1		0	0.97350	1 0	.97084	0.968	001		1	0.93	921	0.953878	k.
			bush	1	0.8	19823	0.9	62663	0.9	73501	L	0 0.	272805	0.217	123	0.86989	4	0.89	318	0.516845	i .
			war		0.8	61211	. 0.9	45219	0.	97084	0.27280	5	0	0 0	.01	0.94273	2	0.963	832	0.75646	<i>i</i> .
			iraq		0.8	69453	0.9	39884	0.9	68001	0.21712	3	0.01		0	0.94762	9	0.940	465	0.732742	1
			louis	iana	0.9	70912	2 0.9	39723	¥	1	0.86989	4 0.	942732	0.947	529		0	0.880	611	0.803738	1
			hurri	cane	. 0.7	24430		1	0	03021	0 8931	8 0	063833	0.940	165	0.88061	1	_	0	0.713962	1
	X	disas	ster	indon	iesia	earth	nquake	bush	010000	war	1rac	0.96045	louis	ana n	JIFFICa	ane flor	da 0.941	3106	962	0	1
	disaster		0165	0 0	.916525		1.911134		062662		0.861211	0.86945	3 (0.970912	0.1	/24439	0.843	3196	_		Ξ.
×	disaste	r i	indone	isia e	arthoua	ke I	hush	t	<u>1.962663</u>		iraq	Louisian:	a	hurricane	flo	rida	b.978	3338			1
disaster			0.0	916525	0.91	1134	0.81	9823	0.861	211	0.869453	0.9	70912	0.72443	9	0.843196	0.95	0/0		<i>: _</i>	·
indonesi	ia 0.9	16525		0	0.012	1	0.96	2663	0.945	5219	0.939884	0.9	39723	0.72110	1	0.978338	0.510	646		. : /	
earthour	ake 0.0	11134		1		0	0.97	3501	0.97	7084	0.968001	0.5	1	0 9397	1	0.953878	0.72	2742		11	
x	disaster	indonesi	a i	earthquake	e bush	1	war		iraq		louisiana	hurricane	e flo	rida	8	0.516845	6.00	7720		: /	
disaster	0	0.91	6525	0.9111	134	0.8198	23 0.	86121	1 0.80	69453	0.970912	0.724	4439	0.843196	2	0.75646	0.003	2062		/.	
indonesia	0.916525		0		1	0.9626	63 0.	94521	9 0.9	39884	0.939723		1	0.978338	5	0.732742	0.713	2002	· /	1.5	
eartnquake	0.911134	0.06	1	0.0720	501	0.9735		27280	+ 0.90	17122	0.960904	0.93	3921	0.953878	1	0.803738		0		STO .	
bush	0.819823	0.96	5210	0.973	084	0 2729	0 0.	2/280	0.2	0.01	0.009894	0.05	3310	0.510045	0	0 713962	_		/ .	~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~	
irao	0.869453	0.94	9884	0.970	001	0.2/20	23	0.0		0.01	0.942732	0.96	1465	0.732742	2	0.715502			· N		
louisiana	0.970912	0.93	9723	0.9000	1	0.8698	94 0.	94273	2 0.94	47629	0.347029	0.880	0611	0.803738	-						
hurricane	0.724439		1	0.939	921	0.893	18 0.	96383	2 0.94	40465	0.880611		0	0.713962							
florida	0.843196	0.97	8338	0.9538	878	0.5168	45 (0.7564	5 0.73	32742	0.803738	0.713	3962	0							
			_			_		_		_											

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).





Outlook

- 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, work flow and result comparison.

What interests Visualizers?

- Need interesting interactions
- Multiple dimensions
- Time depth
- Cross-modular interactions.
- Not just coloring in bits of text that are of interest for linguists.

Summary and Outlook

- Have seen examples of different kinds of visualizations.
- These visualizations allow a new approach to linguistic data.
- Flexible, interactive, make use of the highly skilled human perceptual system.
- More examples to follow tomorrow.
- Now first some design basics.

THANK YOU!

Questions?





http://vis.uni-konstanz.de/