simLC
A Python program for the simulation of language change and language contact

Gerhard Jäger

University of Tübingen
Swedish Collegium for Advanced Study Uppsala

November 5, 2012, Zurich

The Automated Similarity Judgment Program

- Project at MPI EVA in Leipzig around Sören Wichmann
- covers more than 5,000 languages
- basic vocabulary of 40 words for each language, in uniform phonetic transcription
- freely available

used concepts: *I, you, we, one, two, person, fish, dog, louse, tree, leaf, skin, blood, bone, horn, ear, eye, nose, tooth, tongue, knee, hand, breast, liver, drink, see, hear, die, come, sun, star, water, stone, fire, path, mountain, night, full, new, name*

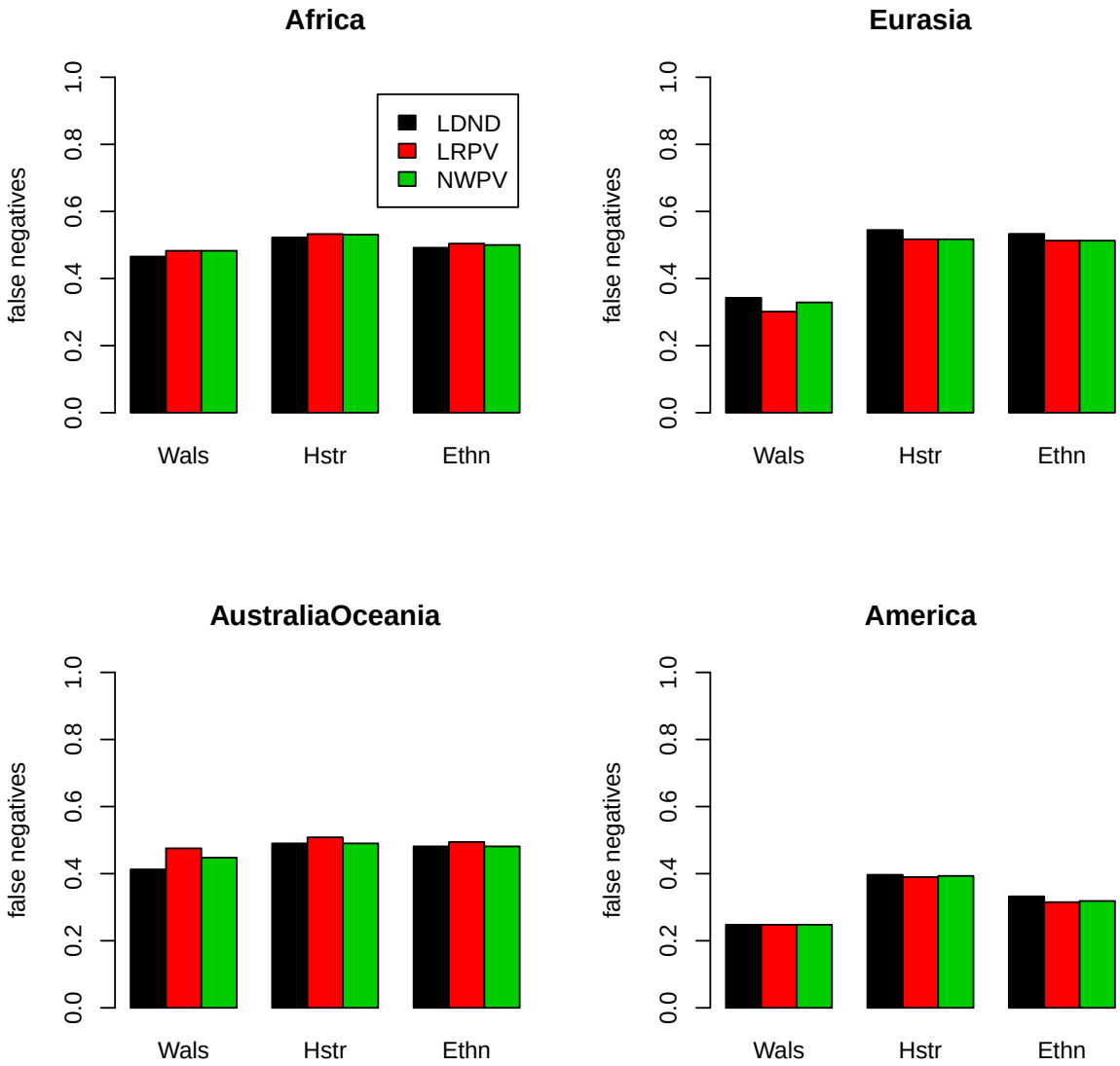
Estimating language similarity

- Basic idea to estimate similarity between two languages (i.e. two Swadesh lists)
  - estimate similarity between translation pairs via some kind of alignment (Levenshtein, weighted alignment, discount vowels, ...)
  - normalize for word length
  - assess probability of false positives, i.e. compare similarities of translation pairs to similarities among unrelated words
- compute distance matrix for all languages of interest and compute a phylogenetic tree using Neighbor Joining (or some other phylogenetic clustering algorithm)
- compare this tree to some expert tree (WALS, Ethnologue)

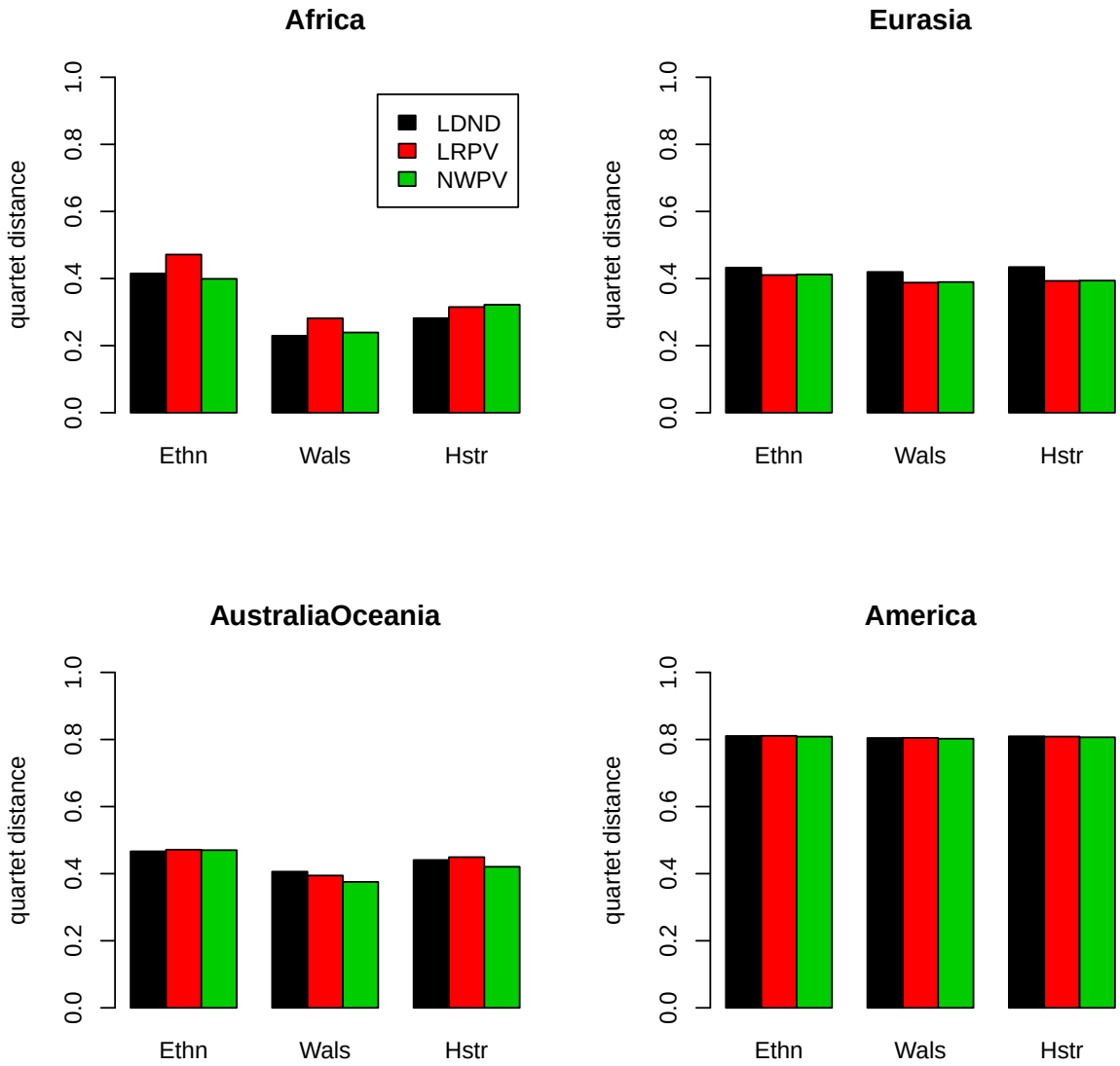
Benchmarking

- comparison of Neighbor-Joining tree with three expert classifications:
  - WALS
  - Ethnologue
  - Hammarström 2010
- measure: proportion of false negatives, i.e. clades in the expert tree that are not recognized in the automatically obtained tree

Benchmarking



Benchmarking



More comparisons

```
Africa
WALS Ethn Hstr
LDND 58 31 490 234 488 248*
LRPV 58 30 490 229 488 242
LR-b 58 30 490 220 488 242
NWPV 58 30 490 230 488 244
NW-b 58 30 490 235* 488 248
LRPV1 58 29 490 220 488 229
LDPV1 58 31 490 235* 488 248*
binom 58 29 490 226 488 238
binom1 58 29 490 229 488 242
binom2 58 33* 490 232 488 244
```

More comparisons

Eurasia

WALS	Ethn		Hstr		
LDND	73	48	325	148	347	162
LRPV	73	51	325	157	347	169
LR-b	73	51	325	156	347	168
NWPV	73	49	325	157	347	169
NW-b	73	48	325	157	347	169
LRPV1	73	51	325	150	347	162
LDPV1	73	50	325	151	347	164
binom	73	52*	325	148	347	162
binom1	73	49	325	149	347	160
binom2	73	52*	325	158*	347	171*

More comparisons

AustraliaOceania

WALS	Ethn	Hstr			
LDND	143	84	610	311	613	318
LRPV	143	75	610	300	613	310
LR-b	143	82	610	299	613	308
NWPV	143	79	610	311	613	318
NW-b	143	81	610	313	613	321
LRPV1	143	78	610	303	613	313
LDPV1	143	87*	610	315*	613	322*
binom	143	87*	610	309	613	321
binom1	143	84	610	312	613	321
binom2	143	83	610	307	613	316

More comparisons

```
America
WALS Ethn Hstr
LDND 109 82 295 178 292 195
LRPV 109 82 295 180 292 200
LR-b 109 82 295 180 292 201
NWPV 109 82 295 179 292 199
NW-b 109 84* 295 181 292 202
LRPV1 109 81 295  173 292  194
LDPV1 109 81 295  179 292  196
binom 109 81 295 182 292 201
binom2 109 82 295 186* 292 205*
binom2 109 81 295 179 292 197
```

More comparisons

	World					
WALS	Ethn		Hstr			
LDND	383	242	1720	871	1740	923
LRPV	383	235	1720	866	1740	921
LR-b	383	245*	1720	855	1740	919
NWPV	383	240	1720	883	1740	935
NW-b	283	239	1720	886*	1749	939*
LRPV1	283	239	1720	848	1749	898
LDPV1	283	245*	1720	880	1749	930
LDPV5	283	245*	1720	878	1749	928
binom	283	245*	1720	865	1749	922
binom1	283	239	1720	875	1749	928
binom2	283	244	1720	876	1749	925

More comparisons

Africa

LDND	0.907	0.844	0.885
LRPV	0.820	0.788	0.844
LR-b	0.790	0.748	0.790
NWPV	0.932*	0.834	0.936*
NW-b	0.844	0.809	0.843
LRPV1	0.886	0.863	0.908
LDPV1	0.932	0.874*	0.900
binom	0.909	0.859	0.902
binom1	0.887	0.856	0.894
binom2	0.904	0.863	0.883

More comparisons

Eurasia

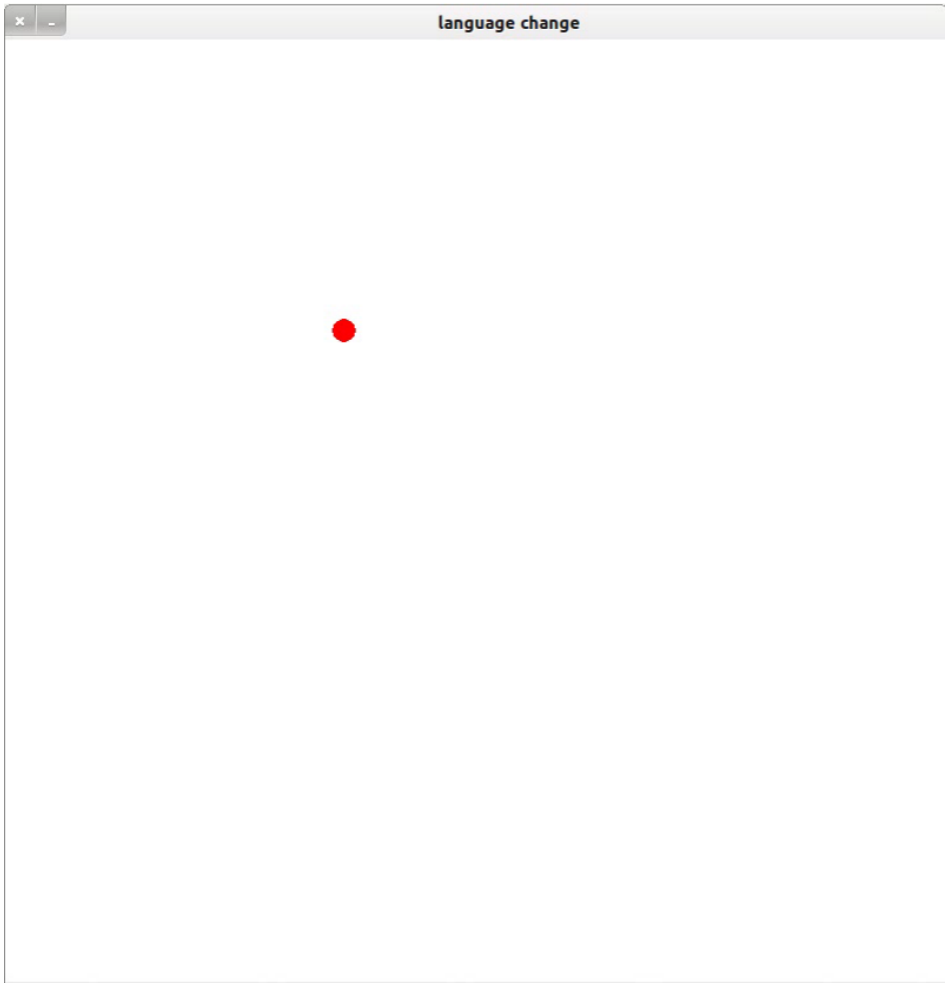
LDND	0.938	0.915	0.915
LRPV	0.974	0.965	0.982
LR-b	0.973	0.963	0.979
NWPV	0.971	0.963	0.980
NW-b	0.970	0.963	0.979
LRPV1	0.966	0.958	0.973
LDPV1	0.982	0.968	0.974
binom	0.983	0.975	0.979
binom1	0.950	0.944	0.959
binom2	0.986*	0.978*	0.983*

More comparisons

	Australia	Oceania	
LDND	0.854	0.763	0.792
LRPV	0.846	0.777	0.780
LR-b	0.875*	0.825*	0.828
NWPV	0.848	0.802	0.821
NW-b	0.865	0.810	0.837*
LRPV1	0.863	0.797	0.835
LDPV1	0.860	0.814	0.835
binom	0.858	0.796	0.832
binom1	0.853	0.783	0.803
binom2	0.833	0.763	0.775

More comparisons

America
LDND 0.889 0.880 0.899
LRPV 0.887 0.878 0.903
LR-b 0.904 0.894 0.921
NWPV 0.898 0.889 0.913
NW-b 0.891 0.880 0.901
LRPV1 0.899 0.890 0.912
LDPV1 0.931 0.920 0.944
binom 0.916 0.904 0.933
binom1 0.940* 0.926* 0.953*
binom2 0.888 0.879 0.901

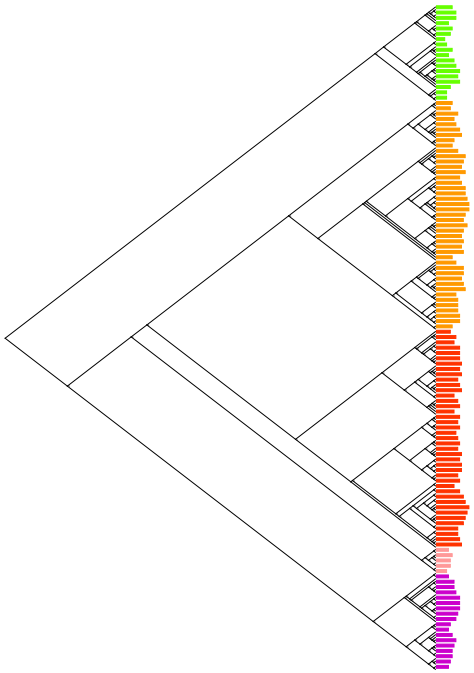

language change

simLC

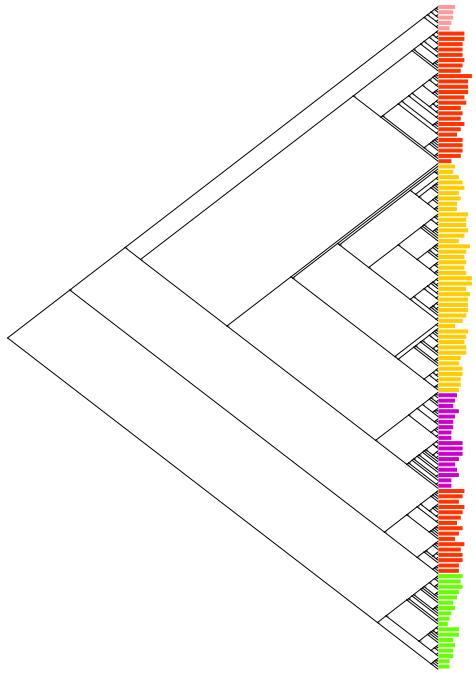
- simulation of language change:
  - "languages" are represented as vectors of identifiers (cognate classes, if you like)
  - languages are located on a two-dimensional surface
  - in each time step, each living language
    - moves a bit around in space
    - may replace words by some new, unrelated words
    - may borrow words from geographically neighboring languages
    - may split into two languages, and
    - may go extinct

Analyzing simulated data

True phylogenetic tree



Reconstructed tree (using neighbor joining)



Analyzing simulated data

Todo: Incorporating sound change

Maximal simplicity

- words are random strings
- sound laws involve replacement of some randomly chosen sound by some other randomly chosen sound

Todo: Incorporating sound change

Maximal realism

- start with phonetically realistic strings (perhaps generated by an HMM trained on real data)
- make sure that length distribution corresponds to empirically obtained distribution
- incorporate differential frequencies of concepts (*I* is more frequent than *horn*, say)
- only use empirically attested sound laws
- perhaps even implement diffusion
- make sure that output of sound shift leads to realistic result (symmetry of phoneme inventory, sufficient discriminability between words)
- chain shifts, context dependency, ...