LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Research Unit
QUANTITATIVE LANGUAGE COMPARISON

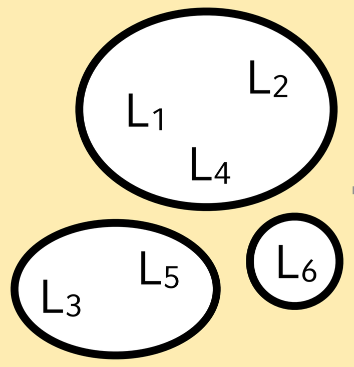
cysouw@lmu.de
MICHAEL CYSOUW

Typology without Types

Languages are different, so shouldn't every language be its own type?

Instead of grouping languages into types, specify the (dis)similarity for every pair of languages

Traditional Typology
Grouping of languages into discrete types



Metrical Representation of Traditional Typology
0 = "same type", 1 = "different type"

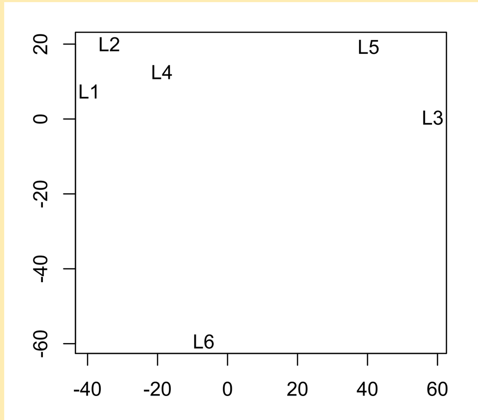
	L_1	L_2	L_3	L_4	L_5	L_6
L_1	0	0	1	0	1	1
L_2	0	0	1	0	1	1
L_3	1	1	0	1	0	1
L_4	0	0	1	0	1	1
L_5	1	1	0	1	0	1
L_6	1	1	1	1	1	0

vs.

Fully specified "Typology without Types"
Values represent pairwise dissimilarity

	L_1	L_2	L_3	L_4	L_5	L_6
L_1	0	31	95	21	87	79
L_2	31	0	98	10	67	83
L_3	95	98	0	79	3	89
L_4	21	10	79	0	50	71
L_5	87	67	3	50	0	90
L_6	79	83	89	71	90	0

Analysis of Structure, Inducing Types
e.g. Multidimensional Scaling



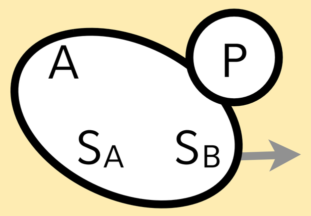
System Typology

Investigate the similarity between constructions within a language

There is no linguistic comparison necessary *between* languages, only *within* languages, which is much easier!
Comparison is purely mathematical. Many typologies can be rephrased in this way, for example alignment:

WITHIN-LANGUAGE ANALYSIS

LANGUAGE I
Nominative Alignment



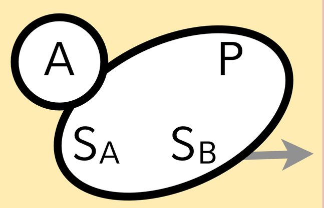
Language-specific Metrical Representation
0 = "same", 1 = "different"

	S_A	S_B	A	P
S_A	0	0	0	1
S_B	0	0	0	1
A	0	0	0	1
P	1	1	1	0

Vector representation of pairwise differences

S_A-S_B	S_A-A	S_A-P	S_B-A	S_B-P	A-P
0	0	1	0	1	1

LANGUAGE II
Ergative Alignment



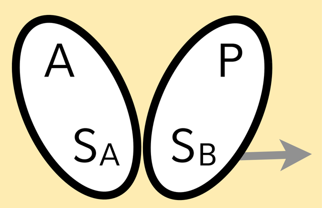
Language-specific Metrical Representation
0 = "same", 1 = "different"

	S_A	S_B	A	P
S_A	0	0	1	0
S_B	0	0	1	0
A	1	1	0	1
P	0	0	1	0

Vector representation of pairwise differences

S_A-S_B	S_A-A	S_A-P	S_B-A	S_B-P	A-P
0	1	0	1	0	1

LANGUAGE III
Agent/Patient Alignment



Language-specific Metrical Representation
0 = "same", 1 = "different"

	S_A	S_B	A	P
S_A	0	1	0	1
S_B	1	0	1	0
A	0	1	0	1
P	1	0	1	0

Vector representation of pairwise differences

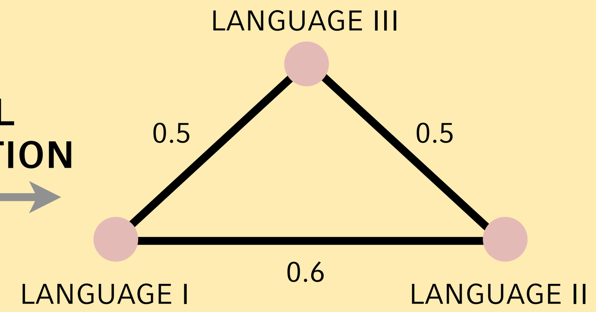
S_A-S_B	S_A-A	S_A-P	S_B-A	S_B-P	A-P
1	0	1	1	0	1

4 out of 6 differences
3 out of 6 differences
3 out of 6 differences

BETWEEN-LANGUAGE COMPARISON

TYPOLOGY	I	II	III
Language I	0	4/6 = 0.6	3/6 = 0.5
Language II	4/6 = 0.6	0	3/6 = 0.5
Language III	3/6 = 0.5	3/6 = 0.5	0

GRAPHICAL REPRESENTATION

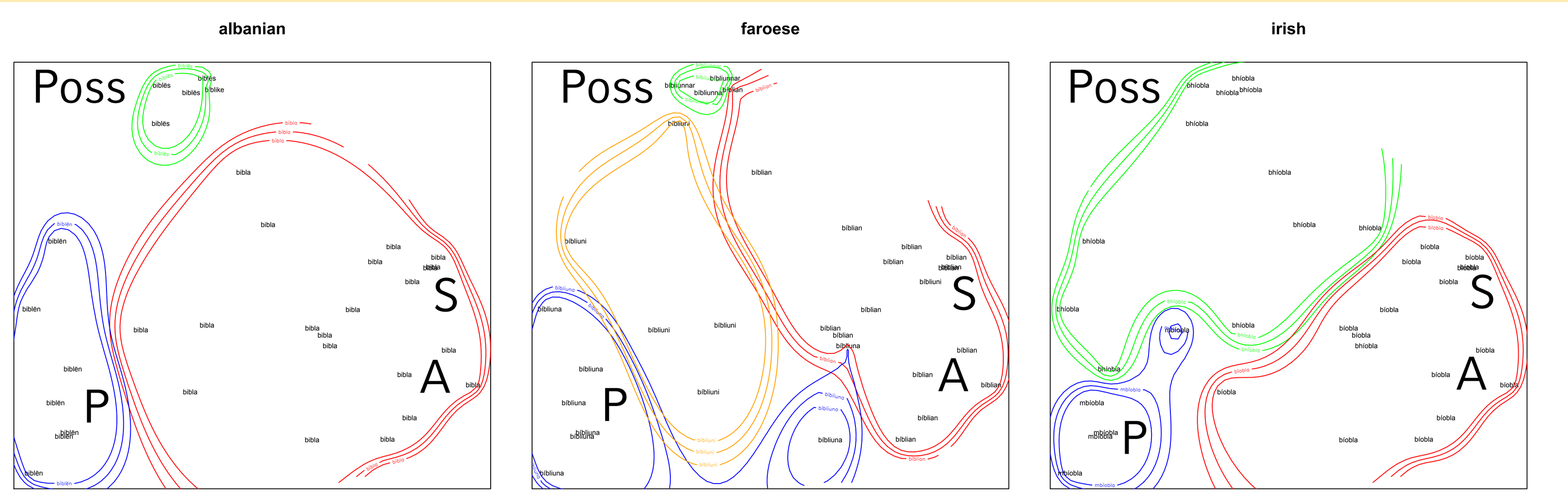


Expanding Typological Detail

With four primitives there are maximally $B_4 = 15$ possible languages, i.e. Bell's Number 4.

To get more detailed representations of the individual structure of each language one can:

- **add more primitives,** e.g. specific verb classes (cf. Figure to the right)
- **use detailed language-specific similarities** constructions in a language are not just identical (0) of different (1), but similar to a certain degree (cf. Numerals, next page)



Alignment as identified by overt case marking in 34 different constructional roles, showing detailed differences and similarities between three nominative languages. Typical S, A, and P roles are indicated. The roles at the top are typical possessors.

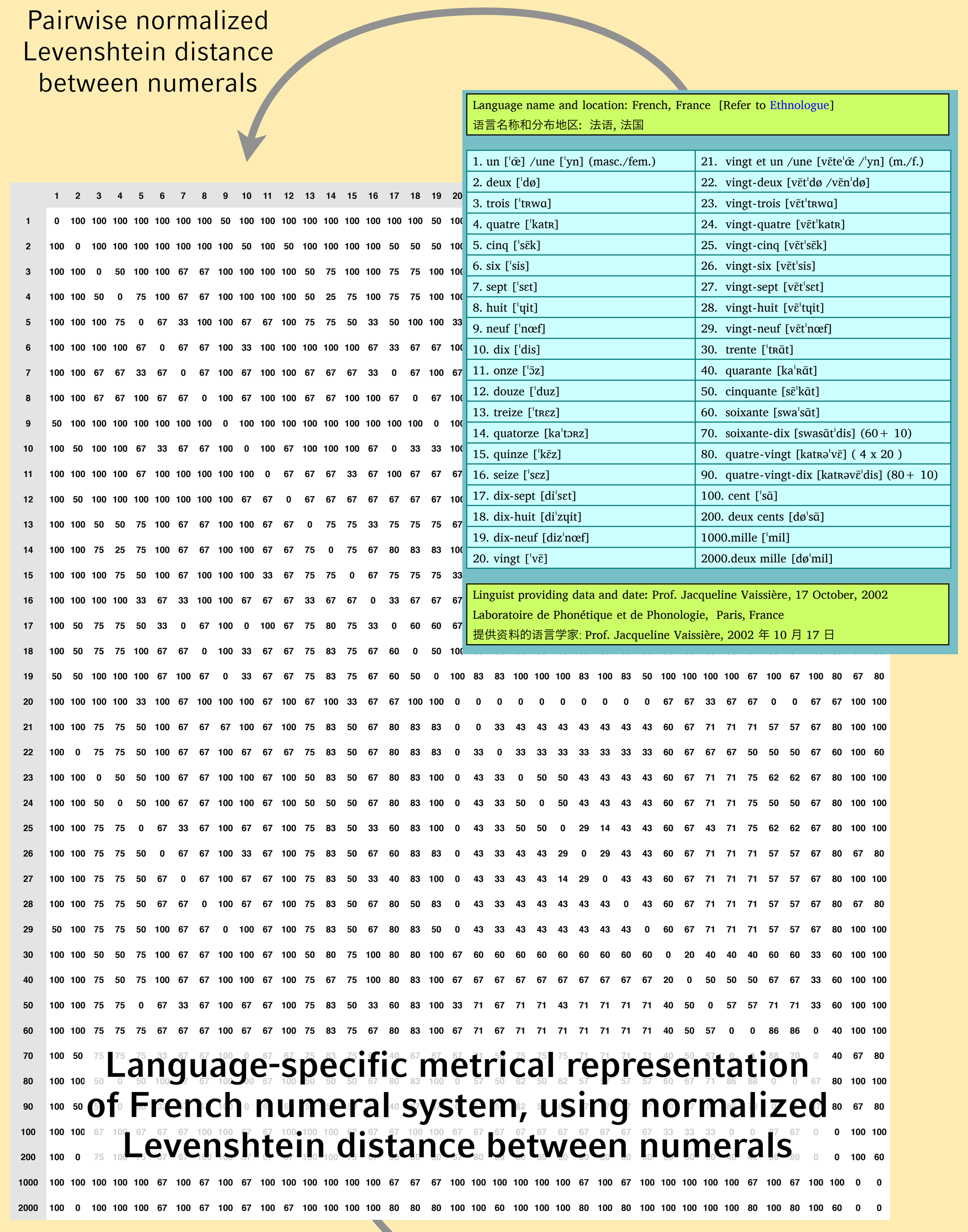
An example of system typology without types

Quantitatively inducing a Numeral System Typology

Based on data collected by
Eugene Chan

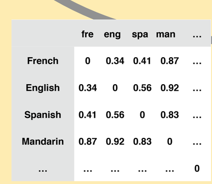
Data preparation with assistance by
Steven Moran & Stefan Koch

Pairwise normalized Levenshtein distance between numerals

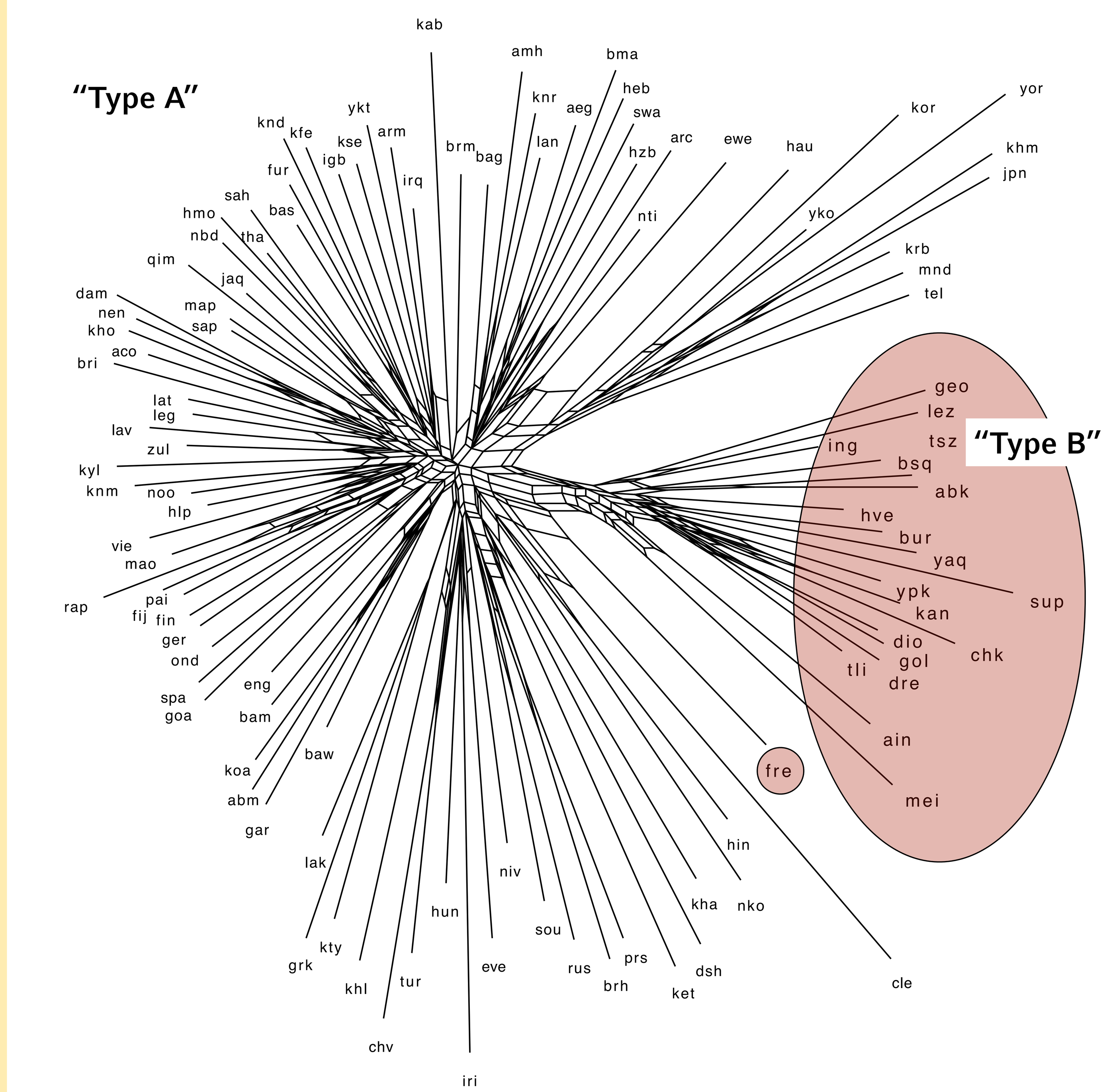


Language-specific metrical representation of French numeral system, using normalized Levenshtein distance between numerals

For every pair of languages, compute Pearson correlation between vector representation of these large matrices



The resulting distance matrix of languages is visualized using SplitsTree



Typology of numeral systems, visualized using a splits graph. There is a split into two groups, with French ("fre") being intermediate.

Interpretation of the Typology

What is the difference between Type A and B ?

- Combine all languages in each group
- Compute the average language-specific metrical specification (i.e. the average of all large matrices as shown above for French)
- This results in an average distance matrix for all numerals for Type A, and one for Type B
- These averages are compared (shown right), showing that for most pairs there is no difference (middle band), but:

The pairs 3-30, 4-40, 5-50, etc. (blue groups) show stronger similarities in Type A, typical of
DECIMAL SYSTEMS

The pairs 2-40, 3-60, 4-80, etc. (red groups) show stronger similarities in Type B, typical of
VIGESIMAL SYSTEMS

Note that there are further subdivisions inside Type A, suggesting that not all decimal systems are equally structured (apparently, some are much more regular than others).