Some Computational Tools for Language Typology

Dik Bakker
Lancaster University
Tools for Typology

**Period 1990 – 2009:**

Computer programs <-> Typological projects
Tools for Typology

Period 1990 – 2009:

Computer programs <-> Typological projects

a. Language sampling
Tools for Typology

Period 1990 – 2009:

Computer programs <-> Typological projects

a. Language sampling
b. Inference of universal implications
Tools for Typology

Period 1990 – 2009:

Computer programs <-> Typological projects

a. Language sampling
b. Inference of universals
c. Lexical classification of languages
Tools for Typology

Period 1990 – 2009:

Computer programs <-> Typological projects

a. Language sampling
b. Inference of universals
c. Lexical classification of languages
d. Language contact and borrowing
Tools for Typology

**Period 1990 – 2009:**

- Computer programs <-> Typological projects
  a. Language sampling
  b. Inference of universals
  c. Lexical classification of languages
  d. Borrowing

...
Tools for Typology

Period 1990 – 2009:

Computer programs <-> Typological projects

a. Language sampling (4)
b. Inference of universals (2)
c. Lexical classification (12)
d. Borrowing (3)
Tools for Typology
Tools for Typology
Tools for Typology

TXT

sampling

inference

classify

CSV

borrow

tools
Tools for Typology

- TXT
- CSV

- sampling
- inference
- classify
- borrow
- tools

Pascal/C++
~
UNIX
Tools for Typology

**Points today:**
Points today:

1. Give an impression of local software (∞)
Tools for Typology

**Points today:**

1. Give an impression of local software (∞)

2. How to make it accessible?
Tools for Typology

Overview:
Tools for Typology

Overview:
1. Sampling
Tools for Typology

Overview:
1. Sampling
2. Inference of universals
Tools for Typology

**Overview:**
1. Sampling
2. Inference of universals
3. Lexical classification
Tools for Typology

Overview:
1. Sampling
2. Inference of universals
3. Lexical classification
4. Nothing about Borrowing:

Tools for Typology

Overview:
1. Sampling
2. Inference of universals
3. Lexical classification
4. Accessibility
1. Language Sampling
Together with:
Kees Hengeveld (Amsterdam)
Peter Kahrel (Amsterdam)
Jan Rijkhoff (Aarhus)

Reference:
'Language sampling'.
Language sampling

Typological project: typically 50 – 500 languages

Question: how to select?
Language sampling

**General issues:**
Language sampling

**General issues:**

- Many features more or less tight to genetic relationships
Language sampling

**General issues:**

- Many features tight to **genetic** relationships

- Areal and contact phenomena
Language sampling

**General issues:**

- Many features tight to genetic relationships
- Areal and contact phenomena
- Distribution of some linguistic features and relations between them are well-known, of (most) others not at all
Language sampling

General issues:

- Many features tight to genetic relationships
- Areal and contact phenomena
- Only some distributions and relations well-known
- Bibliographic gaps
Language sampling

Three types of samples:
Language sampling

Three types of samples:

1. Random sample

→ Only when each language same chance
Language sampling

**Three types of samples:**

1. Random sample

2. Probability sample

→ **Measures chance** on occurrence of certain feature value, or of language type
Language sampling

Three types of samples:

1. Random sample

2. Probability sample

→ Measure chance certain feature value/type

Genetic and areal bias:
  independency ~ (in)stability
Language sampling

Three types of samples:

1. Random sample
2. Probability sample
3. Variety sample

→ Exploration of unknown feature/type: maximum variation
Language sampling

Three types of samples:

1. Random sample → large
2. Probability sample → small
3. Variety sample → intermediate - large
Language sampling

Three types of samples:

1. Random sample → large
2. Probability sample → small
3. Variety sample → intermediate - large
Language sampling

**Variety sample:**

Maximize *variety* $\sim$ maximize *diversity factor*:
Language sampling

Variety sample:

Maximize variety ~ maximize diversity factor:

- language(s) from all families
Language sampling

Variety sample:

Maximize variety $\sim$ maximize diversity factor:

- language(s) from all families

- from as many subgroupings as fit in sample size
DV method

Sample size = n (minimum)

(any language from family for which documentation available)
Sample size > n

(Any language from group for which documentation available)
DV method

Sample size >>> n

(Any language for which documentation available)
DV method

Sample size $\gg n$

(any language for which documentation available)
DV method

Diversity Value per node, based on:
- NOT number of daughter nodes
- NOT on number of daughter languages, but
- Internal Complexity (breadth per level, diminishing)
DV method

Diversity Value per node, based on internal complexity

a. Per family
DV method

Diversity Value per node, based on internal complexity
a. Per family
b. Recursively per lower node
DV method

Procedure:

1. Choose language classification (Ethn/Ruh/Voeg)
DV method

Procedure:
1. Choose language classification (Ethn/Ruh/Voeg)
2. Calculate DV value per node (all tree-like)
DV method

Procedure:

1. Choose language classification
2. Calculate DV value per node
3. Establish sample size (minimum = n of families)
DV method

Procedure:

1. Choose language classification
2. Calculate DV value per node
3. Establish sample size
4. Assign languages to families weighted by DV (> 0)
DV method

Procedure:

1. Choose language classification
2. Calculate DV value per node
3. Establish sample size
4. Assign languages to families weighted by DV
5. Recursively assign languages to lower groups
DV method

Procedure:
1. Choose language classification
2. Calculate DV value per node
3. Establish sample size
4. Assign languages to families weighted by DV
5. Recursively assign languages to lower groups
6. Stop when no languages left to assign
DV method

Procedure:

1. Choose language classification
2. Calculate DV value per node
3. Establish sample size
4. Assign languages to families weighted by DV
5. Recursively assign languages to lower groups
6. Stop when no languages left to assign
7. Optional: select language names (random / criteria)
DV method: results

Classification: Ruhlen91
Criterion 1: Diversity Value
Sample size: 50 (0.95 % of 5273, min=30)
DV method: results

<table>
<thead>
<tr>
<th>Classification</th>
<th>Diversity Value</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afro-Asiatic</td>
<td>55.53/6/258</td>
<td>2</td>
</tr>
<tr>
<td>Altaic</td>
<td>15.07/2/62</td>
<td>1</td>
</tr>
<tr>
<td>Korean-Japanese</td>
<td>3.00/3/4</td>
<td>1</td>
</tr>
<tr>
<td>Australian</td>
<td>67.58/30/262</td>
<td>3</td>
</tr>
<tr>
<td>Austric</td>
<td>137.41/3/1186</td>
<td>5</td>
</tr>
<tr>
<td>Austro-Tai</td>
<td>106.03/2/1027</td>
<td>3</td>
</tr>
<tr>
<td>Austronesian</td>
<td>118.17/4/970</td>
<td>2</td>
</tr>
<tr>
<td>Daic</td>
<td>4.67/2/57</td>
<td>1</td>
</tr>
<tr>
<td>Austroasiatic</td>
<td>28.08/2/155</td>
<td>1</td>
</tr>
<tr>
<td>Miao-Yao</td>
<td>2.00/2/4</td>
<td>1</td>
</tr>
</tbody>
</table>

Sample size: 50 (0.95% of 5273, min=30)
### DV method: results

#### Classification: Ruhlen91

**Criterion 1: Diversity Value**

**Sample size: 50 (0.95 % of 5273, min=30)**

<table>
<thead>
<tr>
<th>Language</th>
<th>Diversity Value</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afro-Asiatic</td>
<td>55.53/6/258</td>
<td>2</td>
</tr>
<tr>
<td>Altaic</td>
<td>15.07/2/62</td>
<td>1</td>
</tr>
<tr>
<td>Korean-Japanese</td>
<td>3.00/3/4</td>
<td>1</td>
</tr>
<tr>
<td>Australian</td>
<td>67.58/30/262</td>
<td>3</td>
</tr>
<tr>
<td>Austric</td>
<td>137.41/3/1186</td>
<td>5</td>
</tr>
<tr>
<td><strong>Austro-Tai</strong></td>
<td><strong>106.03/2/1027</strong></td>
<td>3</td>
</tr>
<tr>
<td>Austronesian</td>
<td>118.17/4/970</td>
<td>2</td>
</tr>
<tr>
<td>Daic</td>
<td>4.67/2/57</td>
<td>1</td>
</tr>
<tr>
<td><strong>Austroasiatic</strong></td>
<td><strong>28.08/2/155</strong></td>
<td>1</td>
</tr>
<tr>
<td>Miao-Yao</td>
<td>2.00/2/4</td>
<td>1</td>
</tr>
</tbody>
</table>

...
### DV method: results

**Classification:** Ruhlen91  
**Criterion 1:** Diversity Value  
**Sample size:** 50 (0.95 % of 5273, min=30)

<table>
<thead>
<tr>
<th>Language</th>
<th>Diversity Value</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afro-Asiatic</td>
<td>55.53/6/258</td>
<td>2</td>
</tr>
<tr>
<td>Altaic</td>
<td>15.07/2/62</td>
<td>1</td>
</tr>
<tr>
<td>Korean-Japanese</td>
<td>3.00/3/4</td>
<td>1</td>
</tr>
<tr>
<td>Australian</td>
<td>67.58/30/262</td>
<td>3</td>
</tr>
<tr>
<td>Austric</td>
<td>137.41/3/1186</td>
<td>5</td>
</tr>
<tr>
<td><strong>Austro-Tai</strong></td>
<td><strong>106.03/2/1027</strong></td>
<td><strong>3</strong></td>
</tr>
<tr>
<td><strong>Austronesian</strong></td>
<td><strong>118.17/4/970</strong></td>
<td><strong>2</strong></td>
</tr>
<tr>
<td><strong>Daic</strong></td>
<td><strong>4.67/2/57</strong></td>
<td><strong>1</strong></td>
</tr>
<tr>
<td><strong>Austroasiatic</strong></td>
<td><strong>28.08/2/155</strong></td>
<td><strong>1</strong></td>
</tr>
<tr>
<td>Miao-Yao</td>
<td>2.00/2/4</td>
<td>1</td>
</tr>
</tbody>
</table>

...
DV method

Options:
DV method

Options:

1. Random selection of languages under nodes
Classification: Ethnologue15
Criterion 1: Diversity Value

Sample size: 150 (2.06% of 7299)

<table>
<thead>
<tr>
<th>Language Group</th>
<th>Diversity Value</th>
<th>Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Austronesian (192.99/12/1268)</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Unclassified (1.00/0/1)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td><strong>1. Ketangalan (G)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>East Formosan (3.00/3/5)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Central (1.00/0/2)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td><strong>2. Amis (G)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bunun (1.00/0/1)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td><strong>3. Bunun (G)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Western Plains (2.00/2/2)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Thao (1.00/0/1)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td><strong>4. Thao (G)</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

...
DV method

Options:

1. Random selection of languages under nodes

2. Stratification on basis of feature values
DV method

Options:

1. Random selection of languages under nodes

2. Stratification on basis of feature values

→ Problem: bibliographic bias
DV method

Options:

1. Random selection of languages under nodes
2. Stratification on basis of feature values
3. Evaluate existing samples
DV method

Options:

1. Random selection of languages under nodes

2. Stratification on basis of feature values

3. Evaluate existing samples

Program has been used for a large number of studies (MA, PhD, articles, books)
2. Inference of Universals
Together with:
Anna Siewierska (Lancaster)

Reference:
UNIVERSALS

Greenberg (1963):
Greenberg (1963):

**Absolute**: Universal 3

Languages with dominant VSO order are *always* prepositional.
Greenberg (1963):  

**Absolute**: Universal 3

Languages with dominant VSO order are *always* prepositional.

**Statistical**: Universal 4

*With overwhelmingly greater than chance frequency,* languages with normal SOV order are postpositional.
LINFER

Data
CSV

inference
LINFER

Data CSV

inference

Result TXT
LINFER

Data
CSV

inference

Result
TXT

<table>
<thead>
<tr>
<th>LGNAME</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abipon</td>
</tr>
<tr>
<td>Abkhaz</td>
</tr>
<tr>
<td>Abun</td>
</tr>
<tr>
<td>Aceh</td>
</tr>
<tr>
<td>Achuma</td>
</tr>
</tbody>
</table>
## LINFER

**Data CSV**

**inference**

**Result TXT**

<table>
<thead>
<tr>
<th>LGNAME</th>
<th>SmrkP</th>
<th>SmrkV</th>
<th>SmrkN</th>
<th>SmrkH</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abipon</td>
<td>123</td>
<td>No</td>
<td>Sgpl</td>
<td>No</td>
</tr>
<tr>
<td>Abkhaz</td>
<td>123</td>
<td>No</td>
<td>Sgpl</td>
<td>No</td>
</tr>
<tr>
<td>Abun</td>
<td>12</td>
<td>No</td>
<td>Sg</td>
<td>No</td>
</tr>
<tr>
<td>Aceh</td>
<td>12</td>
<td>Yes</td>
<td>Nonum</td>
<td>Irr</td>
</tr>
<tr>
<td>Achuma</td>
<td>123</td>
<td>No</td>
<td>Sgdupl</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Tools for Typology
LINFER

Data CSV → inference → Result TXT

<table>
<thead>
<tr>
<th>LGNAME</th>
<th>SmrkP</th>
<th>SmrkV</th>
<th>SmrkN</th>
<th>SmrkH</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abipon</td>
<td>123</td>
<td>No</td>
<td>Sgpl</td>
<td>No</td>
</tr>
<tr>
<td>Abkhaz</td>
<td>123</td>
<td>No</td>
<td>Sgpl</td>
<td>No</td>
</tr>
<tr>
<td>Abun</td>
<td>12</td>
<td>No</td>
<td>Sg</td>
<td>No</td>
</tr>
<tr>
<td>Aceh</td>
<td>12</td>
<td>Yes</td>
<td>Nonum</td>
<td>Irr</td>
</tr>
<tr>
<td>Achuma</td>
<td>123</td>
<td>No</td>
<td>Sgdupl</td>
<td>Yes</td>
</tr>
</tbody>
</table>

SmrkP = 123 → SmrkV = No (ABS)
LINFER

Data CSV

inference

Result TXT

<table>
<thead>
<tr>
<th>LGNAME</th>
<th>SmrkP</th>
<th>SmrkV</th>
<th>SmrkN</th>
<th>SmrkH</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abipon</td>
<td>123</td>
<td>No</td>
<td>Sgpl</td>
<td>No</td>
</tr>
<tr>
<td>Abkhaz</td>
<td>123</td>
<td>No</td>
<td>Sgpl</td>
<td>No</td>
</tr>
<tr>
<td>Abun</td>
<td>12</td>
<td>No</td>
<td>Sg</td>
<td>No</td>
</tr>
<tr>
<td>Aceh</td>
<td>12</td>
<td>Yes</td>
<td>Nonum</td>
<td>Irr</td>
</tr>
<tr>
<td>Achuma</td>
<td>123</td>
<td>No</td>
<td>Sgdupl</td>
<td>Yes</td>
</tr>
</tbody>
</table>

SmrkP = 123  \Rightarrow  SmrkV = No (ABS)

SmrkV = No  \Rightarrow  SmrkP = 123 (0.75)
LINFER

• Automatic inference of implications (A & S):

\[ \Rightarrow \text{generate} + \text{test} \]
LINFER: on WALS


WALS Online: [http://wals.info/](http://wals.info/)

Number of languages: 2558
Number of variables: 143
A first run:

All languages  2558
All variables  139 (minus SignLgs)
LINFER

First run:

All languages  2558
All variables   139

Results:

Potential implications:  413,886
Accepted implications:  1,385 ( = 0.33%)
1. CORSEX=1 <=> CORNUM=1
   0.56 [3] - 0.56 [5]

\[ Fr=0.563, Fa=1.000, Fc=1.000, Fn=1.000, \chi^2<0.5\% \] EQUIV

Sex-based and Non-sex-based Gender System: No gender system
 <=>
Number of Genders: None
1. CORSEX=1, CORNUM=1
   0.56 [3] - 0.56 [5]
   [Fr=0.563, Fa=1.000, Fc=1.000, Fn=1.000, chi2<0.5%] EQUIV

Sex-based and Non-sex-based Gender System: No gender system

<=>
Number of Genders: None

TRIVIAL!!!
15. $KAY_{BCC}=4 \iff VESTAM=2$
\[n=2\]
\[0.13 [7] - 0.13 [4]\]

$[Fr=0.133, Fa=1.000, Fc=1.000, Fn=1.000, \chi^2<0.5\%]$ EQUIV

Number of Basic Colour Categories:
7 or between 7 and 8 categories

$\iff$

Suppletion According to Tense and Aspect:
Suppletion according to aspect
15. $\text{KAYBCC=}4, \text{WESTAM=}2$

0.13 [Fr=0.133, Fa=1.000, Fc=1.000, Fn=1.000, chi2<0.5%] EQUIV

Number of Basic Colour Categories:
7 or between 7 and 8 categories

$\iff$
Suppletion According to Tense and Aspect

Suppletion according to aspect

INSIGNIFICANT!!!
57. \( \text{DRYRPO}=4 \Rightarrow \text{DRYREL}=1 \)
\[ n=291 \]

\[ 0.49 \ [5] - 0.73 \ [7] \]

\[ \text{Fr}=0.486, \text{Fa}=0.983, \text{Fc}=0.655, \text{Fn}=0.511, \text{chi2}<0.5\% \] \ STAT

Relationship between the Order of Object:
Verb-object and prepositional (VO&Prep) 
\[ \Rightarrow \]
Order of Relative Clause and Noun:
Relative clause follows noun (NRel)
57. $\text{DRYRPO}=4 \Rightarrow \text{DRYREL}=1$

$n=291$

$0.49 \ [5] - 0.73 \ [7]$

$[F_r=0.486, \ F_a=0.983, \ F_c=0.655, \ F_n=0.511, \ \chi^2<0.5\%] \ \text{STAT}$

Relationship between the Order of Object:
Verb-object and prepositional (VO&Prep)

$\Rightarrow$

Order of Relative Clause and Noun:
Relative clause follows noun (NRel)
57. \( \text{DRYRPO}=4 \Rightarrow \text{DRYREL}=1 \)
\[ 0.49 \ [5] - 0.73 \ [7] \]
\[ \text{Fr}=0.486, \text{Fa}=0.983, \text{Fc}=0.655, \text{Fn}=0.511, \text{chi}^2<0.5\% \] STAT

Relationship between the Order of Object:
Verb-object and prepositional (VO&Prep)
\[ \Rightarrow \]
Order of Relative Clause and Noun:
Relative clause follows noun (NRel)

\[ \text{EXC: cnt hak mnd squ tuk} \]
LINFER

- Automatic inference of implications (Abs & Stat)
LINFER

• Automatic inference of implications (Abs & Stat)
• Ordered from ‘strongest’ to ‘weakest’
LINFER

- Automatic inference of implications (A & S)
- Ordered from ‘strongest’ to ‘weakest’
- Filtering thresholds
LINFER

- Automatic inference of implications (A & S)
- Ordered from ‘strongest’ to ‘weakest’
- Filtering thresholds
- Selection on subsamples of languages
LINFER

- Automatic inference of implications (A & S)
- Ordered from ‘strongest’ to ‘weakest’
- Filtering thresholds
- Selection on subsamples of languages
- Grouping of variables and values
LINFER

- Automatic inference of implications (A & S)
- Ordered from ‘strongest’ to ‘weakest’
- Filtering thresholds
- Selection on subsamples of languages
- Grouping of variables and values
- Analysis of exceptions
EXPLANATION COUNTEREXAMPLES:

9. BICEXP=5 => NICMTP=1  n= 60
   0.48 [5] - 0.86 [3]

[Fr=0.481, Fa=0.952, Fc=0.531, Fn=0.221, chi2<0.5%] STAT

Exponent of Selected Inflectional Form:
   No case
   =>
M-T Pronouns
   No M-T pronouns

EXC: fre grb lkt
9. BICEXP=5 => NICMTP=1    n= 60

EXC: fre grb lkt

** Possible explaining factors: **
fre:
NICMTP=2 (M-T pronouns, paradigmatic)
HAAEVC=5 (Separate particle)
MADUVU=3 (Uvular continuants only)
grb:
NICMTP=2 (M-T pronouns, paradigmatic)
lkt:
NICMTP=2 (M-T pronouns, paradigmatic)
LINFER

• Automatic inference of implications (A & S)
• Ordered from ‘strongest’ to ‘weakest’
• Filtering thresholds
• Selection on subsamples of languages
• Grouping of variables and values
• Analysis of exceptions
• Chaining of implications (AND/OR)

VO & Prep $\rightarrow$ NRel
Two major questions:
Two major questions:

1. When is an implication statistically reliable?
Two major questions:

1. When is an implication statistically reliable?

2. When is an implication linguistically interesting?
57. DRYRPO=4 => DRYREL=1   n=291

[Fr=0.486, Fa=0.983, Fc=0.655, Fn=0.511, chi2<0.5%] STAT

Relevance: proportion of values for premisse (p / Σpᵢ)
Applicability: proportion of counterexamples (p → ¬ q)
Coverage: proportion of non-premisse languages with conclusion (¬ p → q)
Dominance: proportion of languages with relevant value for variables (p / q)
Negation: proportion of languages with reverse implication (¬ p → ¬ q)
Chi2: for n x m tables (not tetrachoric)
Fisher Exact: when tetrachoric and 1 empty cell
Other statistics: < export data >
57. DRYRPO=4 => DRYREL=1  n=291

[Fr=0.486, Fa=0.983, Fc=0.655, Fn=0.511, \( \chi^2 < 0.5\% \)] STAT

Relevance: proportion of values for premisse (\( p / \sum p_i \))
Applicability: proportion of counterexamples (\( p \rightarrow \neg q \))
Coverage: proportion of non-premisse languages with conclusion (\( \neg p \rightarrow q \))
Dominance: proportion of languages with relevant value for variables (\( p / q \))
Negation: proportion of languages with reverse implication (\( \neg p \rightarrow \neg q \))

\( \chi^2 \): for \( n \times m \) tables (not tetrachoric)
Fisher Exact: when tetrachoric and 1 empty cell

Other statistics: <export data>
57. **DRYRPO=4 => DRYREL=1**  \( n=291 \)

\[ [Fr=0.486, Fa=0.983, Fc=0.655, Fn=0.511, \ \text{chi}2<0.5\%] \] **STAT**

Relevance: proportion of values for premisse \( (p / \Sigma p_i) \)
Applicability: proportion of counterexamples \( (p \rightarrow \neg q) \)
Coverage: proportion of non-premisse languages with conclusion \( (\neg p \rightarrow q) \)
Dominance: proportion of languages with relevant value for variables \( (p / q) \)
Negation: proportion of languages with reverse implication \( (\neg p \rightarrow \neg q) \)
Chi2: for \( n \times m \) tables (not tetrachoric)
Fisher Exact: when tetrachoric and 1 empty cell

**Other statistics:** <export data>
LINFER: DIACHRONY?

**DIACHRONY?**:

99. \( \text{CYSVRB}=3 \Rightarrow \text{CYSIND}=3 \quad n=75 \)

\[
0.40 \ [5] - 0.60 \ [5]
\]

\[
[\text{Fr}=0.395, \text{Fa}=0.949, \text{Fc}=0.625, \text{Fn}=0.628, \chi^2<0.5\%] \quad \text{STAT}
\]

Inclusive/Exclusive Distinction in **Verba**:

- No inclusive/exclusive opposition

=>

Inclusive/Exclusive Distinction in **Pronoun**:

- No inclusive/exclusive opposition

**EXC**: abk cle map mrd
RELATED?:

187. \( \text{HASNPL}=6 \implies \text{BROFIN}=2 \quad n=49 \)

\[
0.49 \ [6] - 0.93 \ [2]
\]

\[\text{Fr}=0.490, \text{Fa}=1.000, \text{Fc}=0.527, \text{Fn}=0.137, \chi^2<1.0\%\] ABS

Occurrence of Nominal Plurality:

\text{Plural in all nouns, always obligatory}

=>

Finger and Hand:

\text{Different words denote 'hand' and 'finger'}
WALS(-like) database, observations:
WALS(-like) database:

- Less than 1:1000 logically possible implications are of potential interest
**WALS(-like) database:**

- Less than 1:1000 logically possible implications are of potential interest
- Most equivalences are trivial
WALS(-like) database:

- Less than 1:1000 logically possible implications are of potential interest
- Most equivalences are trivial
- Many statistically valid implications are hard to interpret linguistically
WALS(-like) database:

• Less than 1:1000 logically possible implications are of potential interest
• Most equivalences are trivial
• Many statistically valid implications are hard to interpret linguistically
• Need for definition: interesting universal
3. Lexical Language Classification
Project ASJP (= Automated Similarity Judgment Program)
ASJP are:

Sören Wichmann (BRD; Netherlands)
Viveka Velupillai (BRD)
André Müller (BRD)
Robert Mailhammer (BRD)
Hagen Jung (BRD)
Eric Holman (USA)
Anthony Grant (UK)
Dmitry Egorov (Russia)
Pamela Brown (USA)
Cecil Brown (USA)
Dik Bakker (UK; Netherlands)
ASJP are:

- Sören Wichmann (BRD; Netherlands)
- Viveka Velupillai (BRD)
- André Müller (BRD)
- Robert Mailhammer (BRD)
- Hagen Jung (BRD)
- Eric Holman (USA)
- Anthony Grant (UK)
- Dmitry Egorov (Russia)
- Pamela Brown (USA)
- Cecil Brown (USA)
- Dik Bakker (UK; Netherlands)
Reference:
ASJP

**Project:**
ASJP (Automated Similarity Judgment Program)

**Overall goal:**
Automatic reconstruction of language relationships
(lexical, grammatical $\rightarrow$ genetic, areal, typological, ...)
**Project:**
ASJP (Automated Similarity Judgment Program)

**Overall goal:**
Automatic reconstruction of language relationships

**Basis:**
Distance matrix between individual languages based on lexical features
**Project:**
ASJP (Automated Similarity Judgment Program)

**Overall goal:**
Automatic reconstruction of language relationships

**Basis:**
Distance matrix between individual languages based on lexical features

**Method:**
Lexicostatistics: mass comparison of *basic* lexical items,
Project: ASJP (Automated Similarity Judgment Program)

Overall goal: Automatic reconstruction of language relationships

Basis: Distance matrix between individual languages based on lexical features

Method: Lexicostatistics: mass comparison of basic lexical items, extended by all relevant data available
Project: ASJP (Automated Similarity Judgment Program)

As in traditional lexicostatistics, but:
ASJP

Project:
ASJP (Automated Similarity Judgment Program)

As in traditional lexicostatistics, but:

1. use of computational algorithms and tools
Project: ASJP (Automated Similarity Judgment Program)

As in traditional lexicostatistics, but:

1. use of computational algorithms and tools
2. methodology from classification in biology
ASJP

Project:
ASJP (Automated Similarity Judgment Program)

WWW
Project: ASJP (Automated Similarity Judgment Program)

WWW

Data sources
ASJP

Project:
ASJP (Automated Similarity Judgment Program)

WWW

Data sources

TOOLS
ASJP

Project:
ASJP (Automated Similarity Judgment Program)

WWW

Data sources

TOOLS

Results

Data Base

Tools for Typology
ASJP Project: ASJP (Automated Similarity Judgment Program)

http://email.eva.mpg.de/~wichmann/ASJPHomePage.htm
ASJP

Project:
ASJP (Automated Similarity Judgment Program)

WWW

Tools for Typology
Overview ASJP system
Overview ASJP system
Overview ASJP system

LEX

ASJP software
Overview ASJP system
Overview ASJP system

<table>
<thead>
<tr>
<th>Language</th>
<th>Language</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>DUTCH</td>
<td>ENGLISH</td>
<td>53.3</td>
</tr>
<tr>
<td>DUTCH</td>
<td>FRENCH</td>
<td>72.7</td>
</tr>
<tr>
<td>DUTCH</td>
<td>MANDARIN</td>
<td>93.8</td>
</tr>
<tr>
<td>…</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Overview ASJP system
Overview ASJP system

Lex

ASJP software

distance matrix

CLASSIF software

Language Classification
Existing Expert Classifications:

- ETHN
- WALS
- EXPRT

**EVALUATION**

- LEX
- ASJP software
- distance matrix
- STAT software
- CLASSIF software

Tools for Typology
Existing Expert Classifications:

- ETHN
- WALS
- EXPRT

EVALUATION

LEX

ASJP software

distance matrix

STAT software

CLASSIF software

Tools for Typology
Tools for Typology

- HIST FACTS
- GEO GRAPH
- ETHN WALS EXPRT
- LEX

ASJP software

distance matrix

MAP software

STAT software

CLASSIF software

Tools for Typology
Lexical items

Data: Word list Morris Swadesh (1955):

100 basic meanings
<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>21. dog</th>
<th>41. nose</th>
<th>61. die</th>
<th>81. smoke</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.</td>
<td>you</td>
<td>22. louse</td>
<td>42. mouth</td>
<td>62. kill</td>
<td>82. fire</td>
</tr>
<tr>
<td>3.</td>
<td>we</td>
<td>23. tree</td>
<td>43. tooth</td>
<td>63. swim</td>
<td>83. ash</td>
</tr>
<tr>
<td>4.</td>
<td>this</td>
<td>24. seed</td>
<td>44. tongue</td>
<td>64. fly</td>
<td>84. burn</td>
</tr>
<tr>
<td>5.</td>
<td>that</td>
<td>25. leaf</td>
<td>45. claw</td>
<td>65. walk</td>
<td>85. path</td>
</tr>
<tr>
<td>6.</td>
<td>who</td>
<td>26. root</td>
<td>46. foot</td>
<td>66. come</td>
<td>86. mountain</td>
</tr>
<tr>
<td>7.</td>
<td>what</td>
<td>27. bark</td>
<td>47. knee</td>
<td>67. lie</td>
<td>87. red</td>
</tr>
<tr>
<td>8.</td>
<td>not</td>
<td>28. skin</td>
<td>48. hand</td>
<td>68. sit</td>
<td>88. green</td>
</tr>
<tr>
<td>9.</td>
<td>all</td>
<td>29. flesh</td>
<td>49. belly</td>
<td>69. stand</td>
<td>89. yellow</td>
</tr>
<tr>
<td>10.</td>
<td>many</td>
<td>30. blood</td>
<td>50. neck</td>
<td>70. give</td>
<td>90. white</td>
</tr>
<tr>
<td>11.</td>
<td>one</td>
<td>31. bone</td>
<td>51. breasts</td>
<td>71. say</td>
<td>91. black</td>
</tr>
<tr>
<td>12.</td>
<td>two</td>
<td>32. grease</td>
<td>52. heart</td>
<td>72. sun</td>
<td>92. night</td>
</tr>
<tr>
<td>13.</td>
<td>big</td>
<td>33. egg</td>
<td>53. liver</td>
<td>73. moon</td>
<td>93. hot</td>
</tr>
<tr>
<td>14.</td>
<td>long</td>
<td>34. horn</td>
<td>54. drink</td>
<td>74. star</td>
<td>94. cold</td>
</tr>
<tr>
<td>15.</td>
<td>small</td>
<td>35. tail</td>
<td>55. eat</td>
<td>75. water</td>
<td>95. full</td>
</tr>
<tr>
<td>16.</td>
<td>woman</td>
<td>36. feather</td>
<td>56. bite</td>
<td>76. rain</td>
<td>96. new</td>
</tr>
<tr>
<td>17.</td>
<td>man</td>
<td>37. hair</td>
<td>57. see</td>
<td>77. stone</td>
<td>97. good</td>
</tr>
<tr>
<td>18.</td>
<td>person</td>
<td>38. head</td>
<td>58. hear</td>
<td>78. sand</td>
<td>98. round</td>
</tr>
<tr>
<td>19.</td>
<td>fish</td>
<td>39. ear</td>
<td>59. know</td>
<td>79. earth</td>
<td>99. dry</td>
</tr>
<tr>
<td>20.</td>
<td>bird</td>
<td>40. eye</td>
<td>60. sleep</td>
<td>80. cloud</td>
<td>100. name</td>
</tr>
</tbody>
</table>
Lexical items: further reduction

Early ASJP analyses have shown:

→ It is not necessary to take all 100 words,
but rather: the MOST STABLE subset
Lexical items: further reduction

Early ASJP analyses have shown:

→ It is not necessary to take all 100 words,
  but rather: the MOST STABLE subset

Least formal variation in accepted classifications

(e.g. Dryer’s Genera; specialized classifications)
<table>
<thead>
<tr>
<th>Language</th>
<th>Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>GERMANIC</td>
<td>FISH</td>
</tr>
<tr>
<td>AFRIKAANS</td>
<td>fis</td>
</tr>
<tr>
<td>BERNESE_GERMAN</td>
<td>fiS</td>
</tr>
<tr>
<td>BRABANTIC</td>
<td>fis</td>
</tr>
<tr>
<td>CIMBRIAN</td>
<td>fiS</td>
</tr>
<tr>
<td>DANISH</td>
<td>fesk</td>
</tr>
<tr>
<td>DUTCH</td>
<td>vis</td>
</tr>
<tr>
<td>ENGLISH</td>
<td>fiS</td>
</tr>
<tr>
<td>FAROESE</td>
<td>fiskur</td>
</tr>
<tr>
<td>FRANS_VLAAMS</td>
<td>fiS</td>
</tr>
<tr>
<td>FRISIAN_WESTERN</td>
<td>fisk</td>
</tr>
<tr>
<td>GOTHIC</td>
<td>fisks</td>
</tr>
<tr>
<td>ICELANDIC</td>
<td>fiskir</td>
</tr>
<tr>
<td>JAMTLANDIC</td>
<td>fisk</td>
</tr>
<tr>
<td>LIMBURGIS</td>
<td>vES</td>
</tr>
<tr>
<td>LUXEMBOURGISI</td>
<td>feS</td>
</tr>
<tr>
<td>NORTH_FRISIAN_AMRUM</td>
<td>fask</td>
</tr>
<tr>
<td>Language</td>
<td>Form</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>-------</td>
</tr>
<tr>
<td>GERMANIC</td>
<td>FISH</td>
</tr>
<tr>
<td>AFRIKAANS</td>
<td>fis</td>
</tr>
<tr>
<td>BERNESE GERMAN</td>
<td>fiS</td>
</tr>
<tr>
<td>BRABANTIC</td>
<td>fis</td>
</tr>
<tr>
<td>CIMBRIAN</td>
<td>fiS</td>
</tr>
<tr>
<td>DANISH</td>
<td>fesk</td>
</tr>
<tr>
<td>DUTCH</td>
<td>vis</td>
</tr>
<tr>
<td>ENGLISH</td>
<td>fiS</td>
</tr>
<tr>
<td>FAROESE</td>
<td>fiskur</td>
</tr>
<tr>
<td>FRANS VLAAMMS</td>
<td>fiS</td>
</tr>
<tr>
<td>FRISIAN WESTERN</td>
<td>fisk</td>
</tr>
<tr>
<td>GOTHIC</td>
<td>fisks</td>
</tr>
<tr>
<td>ICELANDIC</td>
<td>fiskir</td>
</tr>
<tr>
<td>JAMTLANDIC</td>
<td>fisk</td>
</tr>
<tr>
<td>LIMBURGIS</td>
<td>vES</td>
</tr>
<tr>
<td>LUXEMBOURGIS</td>
<td>feS</td>
</tr>
<tr>
<td>NORTH FRISIAN AMRUM</td>
<td>fask</td>
</tr>
<tr>
<td>Language</td>
<td>Sound</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>--------</td>
</tr>
<tr>
<td>GERMANIC</td>
<td>TREE</td>
</tr>
<tr>
<td>AFRIKAANS</td>
<td>bom</td>
</tr>
<tr>
<td>BERNESE_GERMAN</td>
<td>boum</td>
</tr>
<tr>
<td>BRABANTIC</td>
<td>bu3m</td>
</tr>
<tr>
<td>DANISH</td>
<td>trE7</td>
</tr>
<tr>
<td>DUTCH</td>
<td>bom</td>
</tr>
<tr>
<td>ENGLISH</td>
<td>tri</td>
</tr>
<tr>
<td>FAROESE</td>
<td>trEa</td>
</tr>
<tr>
<td>FRANS_VLAAMS</td>
<td>bom</td>
</tr>
<tr>
<td>FRISIAN_WESTERN</td>
<td>bi3m</td>
</tr>
<tr>
<td>GOTHIC</td>
<td>bagms</td>
</tr>
<tr>
<td>ICELANDIC</td>
<td>th<del>ry</del>E</td>
</tr>
<tr>
<td>JAMTLANDIC</td>
<td>tre</td>
</tr>
<tr>
<td>LIMBURGISH</td>
<td>boum</td>
</tr>
<tr>
<td>LUXEMBOURGISH</td>
<td>bam</td>
</tr>
<tr>
<td>NORTH_FRISIAN_AMRUM</td>
<td>bum</td>
</tr>
<tr>
<td>NORTHERN_LOW_SAXON</td>
<td>bom</td>
</tr>
<tr>
<td>NORWEGIAN_BOKMAAL</td>
<td>tre</td>
</tr>
</tbody>
</table>
GERMANIC
AFRIKAANS
BERNESE_GERMAN
BRABANTIC
DANISH
DUTCH
ENGLISH
FAROESE
FRANS_VLAAMS
FRISIAN_WESTERN
GOTHIC
ICELANDIC
JAMTLANDIC
LIMBURGISH
LUXEMBOURGISH
NORTH_FRISIAN_AMRUM
NORTHERN_LOW_SAXON
NORWEGIAN_BOKMAAL

2 forms

Tools for Typology
<table>
<thead>
<tr>
<th>Language</th>
<th>Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>FIN-UGRIC</td>
<td>FISH</td>
</tr>
<tr>
<td>FINNISH</td>
<td>kala</td>
</tr>
<tr>
<td>ESTONIAN</td>
<td>kala</td>
</tr>
<tr>
<td>KARELIAN</td>
<td>kolo</td>
</tr>
<tr>
<td>KILDIN SAAMI</td>
<td>kuly</td>
</tr>
<tr>
<td>KOMI_PERMYAK</td>
<td>Ceri</td>
</tr>
<tr>
<td>KOMI_ZYRIAN</td>
<td>cyeri</td>
</tr>
<tr>
<td>LULE SAAMI</td>
<td>kuole</td>
</tr>
<tr>
<td>MEADOW MARI</td>
<td>kol</td>
</tr>
<tr>
<td>MORDVIN (MOKSHA)</td>
<td>kEl</td>
</tr>
<tr>
<td>NORTH SAAMI</td>
<td>guoli</td>
</tr>
<tr>
<td>SKOLT SAAMI</td>
<td>kuel</td>
</tr>
<tr>
<td>SOUTH SAAMI</td>
<td>gueli3</td>
</tr>
<tr>
<td>UDMURT</td>
<td>cyorig</td>
</tr>
<tr>
<td>VEPS</td>
<td>kala</td>
</tr>
<tr>
<td>NENETS</td>
<td>xaly</td>
</tr>
<tr>
<td>SELKUP</td>
<td>q3l3</td>
</tr>
<tr>
<td>CSANGO</td>
<td>hol</td>
</tr>
<tr>
<td>HUNGARIAN</td>
<td>hal</td>
</tr>
</tbody>
</table>
FIN-UGRIC  FISH
FINNISH    kala
ESTONIAN   kala
KARELIAN   kolo
KILDIN_SAAMI kuly
KOMI_PERMYAK Ceri
KOMI_ZYRIAN cyeri
LULE_SAAMI kuole
MEADOW_MARI kol
MORDVIN(MOKSHA) kEl
NORTH_SAAMI guoli
SKOLT_SAAMI kuel
SOUTH_SAAMI gueli3
UDMURT cyorig
VEPS kala
NENETS xaly
SELKUP q3I3
CSANGO hol
HUNGARIAN hal

1 proto form
FIN-UGRIC TREE
FINNISH puu
INARI_SAAMI muoro
KARELIAN pu
KILDIN_SAAMI mur
KOMI_PERMYAK pu
KOMI_ZYRIAN pu
LULE_SAAMI muora
MEADOW_MARI puSeNxe
MORDVIN(MOKSHA) SuftE
NORTH_SAAMI muoro
SKOLT_SAAMI mu3r | mw3r
SOUTH_SAAMI moer3
UDMURT pispu
VEPS pu
NENETS pya
SELKUP po
CSANGO fo
HUNGARIAN fa
Lexical items: further reduction

Early analyses have shown:

Most stable 40/100 item subset gives:
Lexical items: further reduction

Early analyses have shown:

Most stable 40/100 item subset gives:

- at least the same results as > 40
Early analyses have shown:

Most stable 40/100 item subset gives:

- at least the same results as $> 40$
- better results than $< 40$
Correlation vs. Number of items

Ethnologue Classification

WALS Classification
The diagram compares the correlation between Ethnologue Classification and WALS Classification. The y-axis represents the correlation values, ranging from 0 to 1, and the x-axis shows the number of items, ranging from 0 to 100. The graph shows the trend of correlation as the number of items increases.
<table>
<thead>
<tr>
<th>I</th>
<th>dog</th>
<th>nose</th>
<th>die</th>
<th>smoke</th>
</tr>
</thead>
<tbody>
<tr>
<td>you</td>
<td>louse</td>
<td>mouth</td>
<td>kill</td>
<td>fire</td>
</tr>
<tr>
<td>we</td>
<td>tree</td>
<td>tooth</td>
<td>swim</td>
<td>ash</td>
</tr>
<tr>
<td>this</td>
<td>seed</td>
<td>tongue</td>
<td>fly</td>
<td>burn</td>
</tr>
<tr>
<td>that</td>
<td>leaf</td>
<td>claw</td>
<td>walk</td>
<td>path</td>
</tr>
<tr>
<td>who</td>
<td>root</td>
<td>foot</td>
<td>come</td>
<td>mountain</td>
</tr>
<tr>
<td>what</td>
<td>bark</td>
<td>knee</td>
<td>lie</td>
<td>red</td>
</tr>
<tr>
<td>not</td>
<td>skin</td>
<td>hand</td>
<td>sit</td>
<td>green</td>
</tr>
<tr>
<td>all</td>
<td>flesh</td>
<td>belly</td>
<td>stand</td>
<td>yellow</td>
</tr>
<tr>
<td>many</td>
<td>blood</td>
<td>neck</td>
<td>give</td>
<td>white</td>
</tr>
<tr>
<td>one</td>
<td>bone</td>
<td>breast</td>
<td>say</td>
<td>black</td>
</tr>
<tr>
<td>two</td>
<td>grease</td>
<td>heart</td>
<td>sun</td>
<td>night</td>
</tr>
<tr>
<td>big</td>
<td>egg</td>
<td>liver</td>
<td>moon</td>
<td>hot</td>
</tr>
<tr>
<td>long</td>
<td>horn</td>
<td>drink</td>
<td>star</td>
<td>cold</td>
</tr>
<tr>
<td>small</td>
<td>tail</td>
<td>eat</td>
<td>water</td>
<td>full</td>
</tr>
<tr>
<td>woman</td>
<td>feather</td>
<td>bite</td>
<td>rain</td>
<td>new</td>
</tr>
<tr>
<td>man</td>
<td>hair</td>
<td>see</td>
<td>stone</td>
<td>good</td>
</tr>
<tr>
<td>person</td>
<td>head</td>
<td>hear</td>
<td>sand</td>
<td>round</td>
</tr>
<tr>
<td>fish</td>
<td>ear</td>
<td>know</td>
<td>earth</td>
<td>dry</td>
</tr>
<tr>
<td>bird</td>
<td>eye</td>
<td>sleep</td>
<td>cloud</td>
<td>name</td>
</tr>
<tr>
<td></td>
<td>dog</td>
<td>nose</td>
<td>die</td>
<td>smoke</td>
</tr>
<tr>
<td>----</td>
<td>-----</td>
<td>------</td>
<td>-----</td>
<td>-------</td>
</tr>
<tr>
<td>you</td>
<td>louse</td>
<td>mouth</td>
<td>kill</td>
<td>fire</td>
</tr>
<tr>
<td>we</td>
<td>tree</td>
<td>tooth</td>
<td>swim</td>
<td>ash</td>
</tr>
<tr>
<td>this</td>
<td>seed</td>
<td>tongue</td>
<td>fly</td>
<td>burn</td>
</tr>
<tr>
<td>that</td>
<td>leaf</td>
<td>claw</td>
<td>walk</td>
<td>path</td>
</tr>
<tr>
<td>who</td>
<td>root</td>
<td>foot</td>
<td>come</td>
<td>mountain</td>
</tr>
<tr>
<td>what</td>
<td>bark</td>
<td>knee</td>
<td>lie</td>
<td>red</td>
</tr>
<tr>
<td>not</td>
<td>skin</td>
<td>hand</td>
<td>sit</td>
<td>green</td>
</tr>
<tr>
<td>all</td>
<td>flesh</td>
<td>belly</td>
<td>stand</td>
<td>yellow</td>
</tr>
<tr>
<td>many</td>
<td>blood</td>
<td>neck</td>
<td>give</td>
<td>white</td>
</tr>
<tr>
<td>one</td>
<td>bone</td>
<td>breast</td>
<td>say</td>
<td>black</td>
</tr>
<tr>
<td>two</td>
<td>grease</td>
<td>heart</td>
<td>sun</td>
<td>night</td>
</tr>
<tr>
<td>big</td>
<td>egg</td>
<td>liver</td>
<td>moon</td>
<td>hot</td>
</tr>
<tr>
<td>long</td>
<td>horn</td>
<td>drink</td>
<td>star</td>
<td>cold</td>
</tr>
<tr>
<td>small</td>
<td>tail</td>
<td>eat</td>
<td>water</td>
<td>full</td>
</tr>
<tr>
<td>woman</td>
<td>feather</td>
<td>bite</td>
<td>rain</td>
<td>new</td>
</tr>
<tr>
<td>man</td>
<td>hair</td>
<td>see</td>
<td>stone</td>
<td>good</td>
</tr>
<tr>
<td>person</td>
<td>head</td>
<td>hear</td>
<td>sand</td>
<td>round</td>
</tr>
<tr>
<td>fish</td>
<td>ear</td>
<td>know</td>
<td>earth</td>
<td>dry</td>
</tr>
<tr>
<td>bird</td>
<td>eye</td>
<td>sleep</td>
<td>cloud</td>
<td>name</td>
</tr>
</tbody>
</table>
Lexical items: further reduction

Early analyses have shown:

- Most stable 40/100 item subset gives optimal results

→ Less work
Lexical items: further reduction

Early analyses have shown:

- Most stable 40/100 item subset gives optimal results

→ Less work

→ Less missing data
Lexical items: further reduction

Early analyses have shown:

- Most stable 40/100 item subset gives optimal results
  → Less work
  → Less missing data
  → Faster processing; combinatorial explosion:

\[
\frac{40}{100} \sim 2.5 \times 2.5 = 6.3
\]
Current sample

3500 languages * 40 lexical items
Languages currently sampled
Processing problems ...

3500 languages * 40 lexical items:

~ 10,000,000,000 comparisons .....
Processing problems ...

3500 languages * 40 lexical items:

~ 10.000.000.000 comparisons ..... (10G)

→ comparison at the phoneme level

for feature level: ~ 250.000.000.000 (0.25T)
Processing problems ...
Solution: parallel processing

100 times faster
Lexical items: transcription

First phase of project (2007):

Problems with full phonological (IPA) representation of words:
Lexical items: transcription

First phase of project (2007):

Problems with full IPA representation of words:

- data entry via keyboard
Lexical items: transcription

First phase of project (2007):

Problems with full IPA representation of words:

- data entry via keyboard

- simple programming languages (Fortran; Pascal)
Lexical items: transcription

First phase of project (2007):

Problems with full IPA representation of words:

- data entry via keyboard

- simple programming languages (Fortran; Pascal)

→ Recoding to simplified ASJPcode (keyboard)
Lexical items: transcription

**ASJPcode:**

7 Vowels

34 Consonants

All other phonemes to ‘closest sound’
Lexical items: transcription

**ASJPcode:**

- 7 Vowels
- 34 Consonants

**Symbols for:**

- Nasalization
- Labialization
- Palatalization
- Aspiration
- Glottalization

All other phonemes to ‘closest sound’
Abaza (Caucasian):

**Meaning**

PERSON

LEAF

SKIN

HORN

NOSE

TOOTH
### Abaza (Caucasian):

<table>
<thead>
<tr>
<th>Part</th>
<th>Meaning</th>
<th>IPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>PERSON</td>
<td>埱ncoder</td>
<td>ꝩvĩʈʃ’jw&gt;(),wĩs</td>
</tr>
<tr>
<td>LEAF</td>
<td>埱onk</td>
<td>bγj+i</td>
</tr>
<tr>
<td>SKIN</td>
<td>埱onk</td>
<td>tʃwazj</td>
</tr>
<tr>
<td>HORN</td>
<td>埱onk</td>
<td>tʃ’wĩꜩw’a</td>
</tr>
<tr>
<td>NOSE</td>
<td>埱onk</td>
<td>pɪnts’ä</td>
</tr>
<tr>
<td>TOOTH</td>
<td>埱onk</td>
<td>pɪts</td>
</tr>
</tbody>
</table>
Abaza (Caucasian):

<table>
<thead>
<tr>
<th>Meaning</th>
<th>IPA</th>
<th>ASJPcode</th>
</tr>
</thead>
<tbody>
<tr>
<td>PERSON</td>
<td>ŋw̞iʧ̞'jw̞iʃ̞</td>
<td>Xw3Cw&quot;yXw3s</td>
</tr>
<tr>
<td>LEAF</td>
<td>bγ̊j̞i</td>
<td>bxy3</td>
</tr>
<tr>
<td>SKIN</td>
<td>ŋj̞w̞az̞j</td>
<td>Cwazy</td>
</tr>
<tr>
<td>HORN</td>
<td>ŋw̞iʧ̞'wa</td>
<td>Cw&quot;3Xwa</td>
</tr>
<tr>
<td>NOSE</td>
<td>pints'a</td>
<td>p3nc&quot;a</td>
</tr>
<tr>
<td>TOOTH</td>
<td>pỊts</td>
<td>p3c</td>
</tr>
</tbody>
</table>
Loss of information?

Experiment with Caucasian (39 lgs):
Loss of information?

**Experiment with Caucasian (39 lgs):**

- Full IPA does not score better for separating language families
Loss of information?

Experiment with Caucasian (39 lgs):

- Full IPA does not score better for separating language families

- For *precise genetic classification* IPA is even less accurate than ASJP code (too specific?)
Comparing words

**Most important measure:** Levenshtein Distance
Comparing words

Levenshtein Distance (LD)

a. between 2 words:
Comparing words

**Levenshtein Distance (LD)**

a. **between 2 words:**

number of transformations (=changes & additions) to get from the shorter form to the longer one
Comparing words

Levenshtein Distance (LD)

a. between 2 words:

number of transformations (=changes & additions) to get from the shorter form to the longer one

b. between 2 languages:

mean LD for all common pairs
Comparing words

Two problems with simple LD:
Comparing words

Two problems:

1. Value depends on length of longest word
Comparing words

1. Value depends on length of longest word

CAT

DOG

x x x = 3
Comparing words

1. Value depends on length of longest word

CAT

DOG

XXX = 3

ELEPHANT

DOG

XXXXXXXXXX = 8
Comparing words

1. Value depends on length of longest word

Normalize: \( \text{LDN} = \left( \frac{\text{LD}}{L_{\text{max}}} \right) \)
Comparing words

1. Value depends on length of longest word

CAT
DOG
x x x = 3/3 = 1.0

ELEPHANT
DOG
x x x x x x x x = 8 / 8 = 1.0
Comparing words

Two problems:

1. Value depends on length of longest word

   → Normalize: \( \text{LDN} = ( \frac{\text{LD}}{L_{\text{max}}} ) \)

2. Differences between lgs in phonological overlap
Comparing words

2. Differences between lgs in phonological overlap

DUTCH ~ ENGLISH: mean LDN: 0.55
Comparing words

2. Differences between lgs in phonological overlap

DUTCH ~ ENGLISH: mean LDN: 0.55

DUTCH ~ MANDARIN: mean LDN: 0.91
2. Differences between lgs in phonological overlap

DUT ~ ENG:  mean LDN: 0.55

mean LDN other words: 0.89

DUT ~ MAN:  mean LDN: 0.91

mean LDN other words: 0.93
Comparing words

2. Differences between lgs in phonological overlap

DUT ~ ENG: mean LDN: 0.55 / 0.89
mean LDN other words: 0.89

DUT ~ MAN: mean LDN: 0.91 / 0.93
mean LDN other words: 0.93
Comparing words

2. Differences between lgs in phonological overlap

DUT ~ ENG: mean LDN: 0.55 / 0.89 = 0.62

DUT ~ MAN: mean LDN: 0.91 / 0.93 = 0.99
Comparing words

Two problems:

1. Value depends on length of longest word

   → Normalize: \( \text{LDN} = \left( \frac{\text{LD}}{L_{\text{max}}} \right) \)

2. Differences between lgs in phonological overlap

   → Eliminate ‘background noise’:

   \( \text{LDND} = \left( \frac{\text{LDN}}{\text{LDN}_{\text{different pairs}}} \right) \)
Sino-Tibetan: Chinese

ASJP tree based on lexical relations
Sino-Tibetan: Chinese

ASJP tree based on lexical relations

Middle Chinese
  Old Chinese
    Suzhou Wu
      Amoy Minnan
        Hainan Minnan
          Mandarin
            Cantonese
              Hakka

ALL & ONLY
Sino-Tibetan: Chinese

Genetic classification in Thurgood & LaPolla (eds)
Lexical plus typological data

Swadesh (3500) ~ WALS (2580)

ASJP

distance matrices

TREE SFTW

Tools for Typology
'SWALSH'

ASJP

distance matrices

TREE

SFTW

Tools for Typology
Improving the fit

Only WALS

Etnologue 2005

Tools for Typology

Percentage LDND

Correlations

Only ASJP
Improving the fit

Correlation

0.25000

0.20000

0.17500

0
25
50
75
100

Percentage LDND

Only WALS

Only ASJP

Etnologue 2005

ONLY 550 LANGUAGES !!!
Lexical items: transcription

Second phase of project (2009-10):

Replace ASJP code by full IPA representations
Lexical items: transcription

Second phase of project (2009-10):

Problems with full IPA representation solved:
Second phase of project (2009-10):

Problems with full IPA representation solved:

1. scan/download/... full IPA representations
Lexical items: transcription

**Second phase of project (2009-10):**

Problems with full IPA representation solved:

1. scan/download/… full IPA representations

2. automatic conversion IPA to integer (Python)
Lexical items: transcription

Second phase of project (2009-10):

Problems with full IPA representation solved:

1. scan/download/… full IPA representations

2. automatic conversion IPA to integer (Python)

3. (semi-)automatic recoding to ASJPcode: transduction on the basis of a formal grammar
Lexical items: transcription

Abaza (Caucasian):
Meaning: PERSON
Lexical items: transcription

Abaza (Caucasian):
Meaning: PERSON

IPA: ʃwɪʧ’jwʃw’ɪʃ
Lexical items: transcription

Abaza (Caucasian):
Meaning: PERSON

IPA: ʒwิtʃ' jwɔwɔs

Decimal: 661,695,616,679,700,690,695,661,695,616,115
Lexical items: transcription

Abaza (Caucasian):
Meaning: PERSON

IPA: ʒwɪʧ'jwɔwɪs

Decimal: 661,695,616,679,700,690,695,661,695,616,115

'а' <- 661, 895, 416, ...

formal grammar
Lexical items: transcription

Abaza (Caucasian):
Meaning: PERSON

IPA: ʧwɪʃˈjwɔʃis

Decimal: 661,695,616,679,700,690,695,661,695,616,115

'a' <- 661, 895, 416, ...

formal grammar

ASJP++code
Lexical items: transcription

IPA: \( \text{'\textit{\textbf{w}+\textit{\textbf{i}t}\text{'}\text{-}\text{\textit{\textbf{jw}w}i}s}} \)

Decimal: 661,695,616,679,700,690,695,661,695,616,115

‘a’ <- 661, 895, 416, ...

... ‘a’ [+Vow, +Low, +Middle]

‘b’ [+Cons, +Labial, +Plosive, +Voice ]

formal grammar + phonological features

ASJP++code: ( comparison of phonological features )
5. Accessibility
Accessibility
Small Tools:
- Lookup Ethnologue code
- Affiliation
- Linguistic variables
- ...
Access: data, internal

Generally accepted data structures (Unicode; UTF8)
Access: data, internal

Generally accepted data structures (Unicode; UTF8)
Access: data, internal

Tools for Typology
Access: data, external

Tools for Typology
Access: data, universal?
Access: software: in/external?

1. Too big / slow for Windows (?)
Access: software: in/external?

1. Too big / slow for Windows (?)
2. No user interface
Access: software: in/external?

1. Too big / slow for Windows (?)
2. No user interface
Access: software: in/external?

1. Too big / slow for Windows (?)
2. No user interface
Access: software: in/external?

1. Too big / slow for Windows (?)  
2. No user interface
There must be more of such out there, some *useful* for the linguistic community, but:
Accessibility requirements

a. platform
   - accessible from WWW
   - programming language
Accessibility requirements

a. platform
   - accessible from WWW
   - programming language

b. ‘human’ interface
   - interactive interface < - > actual application
   - user documentation
Accessibility requirements

a. platform
   - accessible from WWW
   - programming language

b. ‘human’ interface
   - interactive interface < - > actual application
   - user documentation

c. data structure
   - TXT, CSV → HTML, Java Script, ... (??)
Accessibility requirements

a. platform
   - accessible from WWW
   - programming language

b. ‘human’ interface
   - interactive interface < - > actual application
   - user documentation

c. data structure
   - TXT, CSV → HTML, Java Script, … (?)

d. maintenance
   - programmer documentation