Some Computational Tools for Language Typology

Dik Bakker

Lancaster University

Tools for Typology

<u>Period 1990 – 2009:</u>

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

...

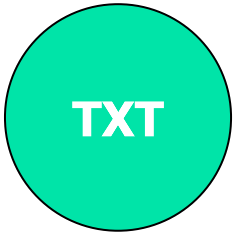
Tools for Typology

Period 1990 – 2009:

Computer programs <-> Typological projects

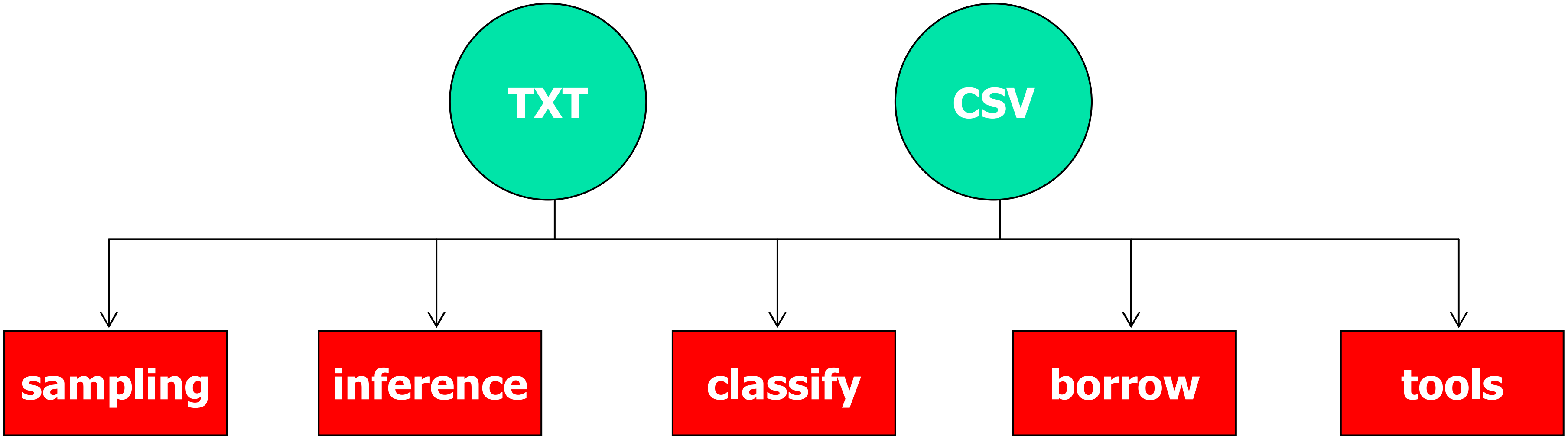
a. Language sampling (4)
b. Inference of universals (2)
c. Lexical classification (12)
d. Borowing (3)

Tools for Typology

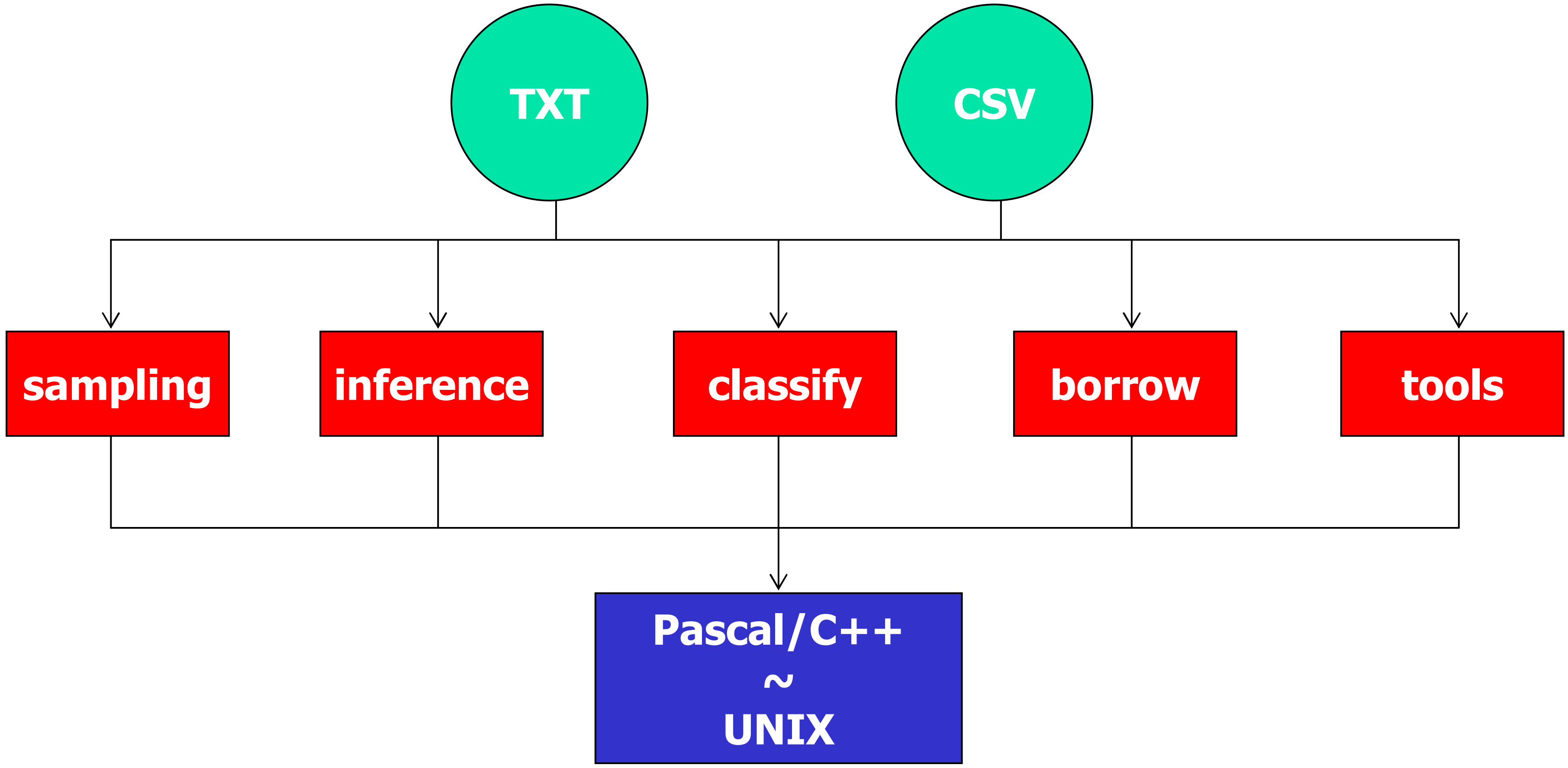


Tools for Typology

Tools for Typology



Tools for Typology



Tools for Typology

Points today:

Tools for Typology

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

<u>Overview:</u>

1. Sampling

2. Inference of universals

3. Lexical classification

4. Nothing about Borrowing:

Bakker, D., J. Gómez-Rendón & E. Hekking (2008).
'Spanish meets Guaraní, Otomí and Quichua: a multilingual confrontation'.
In Th. Stolz, D. Bakker & R. Palomo (eds)
***Aspects of Language Contact*. Mouton de Gruyter, 165-238.**

Tools for Typology

<u>Overview:</u>

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:
Rijkhoff J. & D. Bakker (1998). 'Language sampling'. *Linguistic Typology* 2-3, 263-314.

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

<u>Three types of samples:</u>

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 ~ maximize diversity factor:

Language sampling

Variety sample:

Maximize variety ~ maximize diversity factor:

- **language(s) from all families**

Language sampling

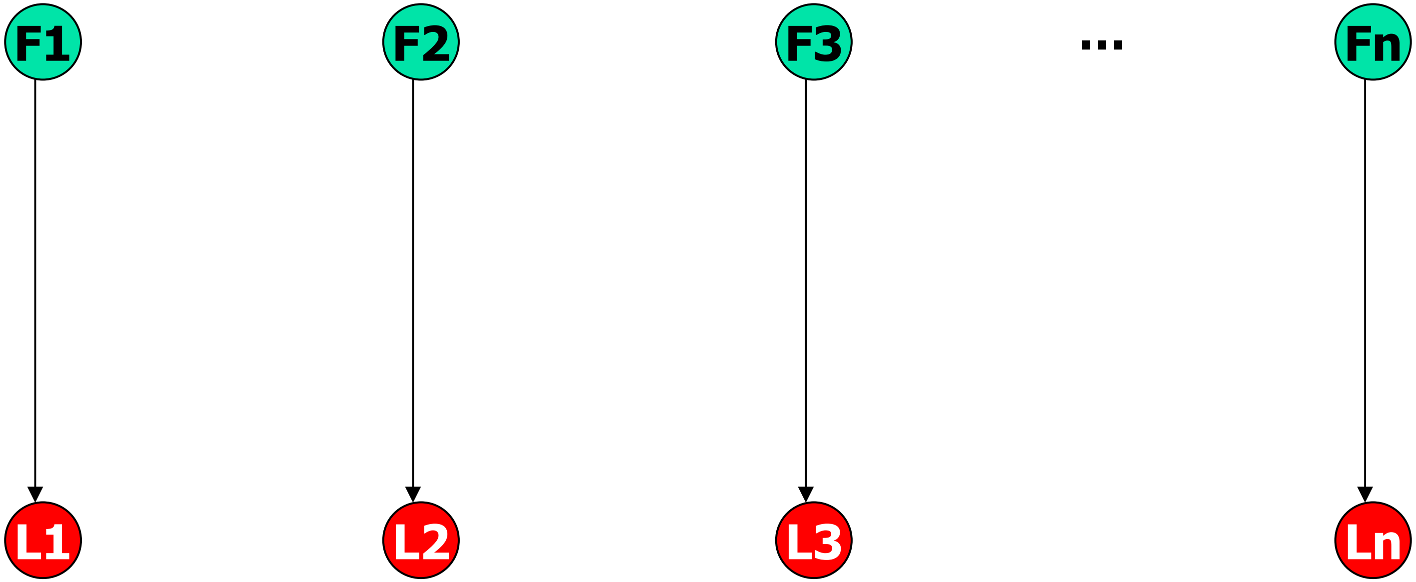
Variety sample:

Maximize variety ~ maximize diversity factor:

- **language(s) from all families**
- **from as many subgroupings as fit in sample size**

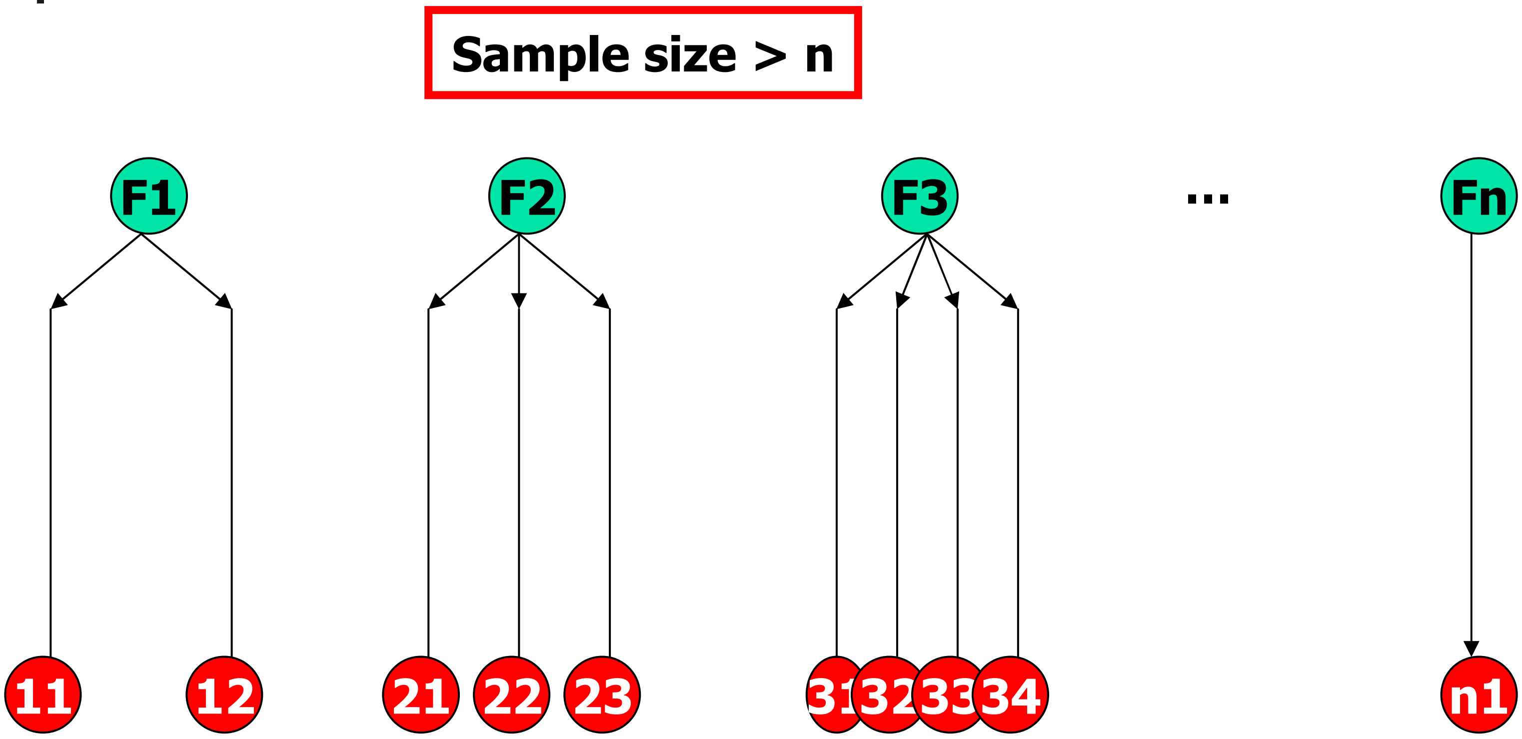
DV method

Sample size = n (minimum)



(any language from family for which documentation available)

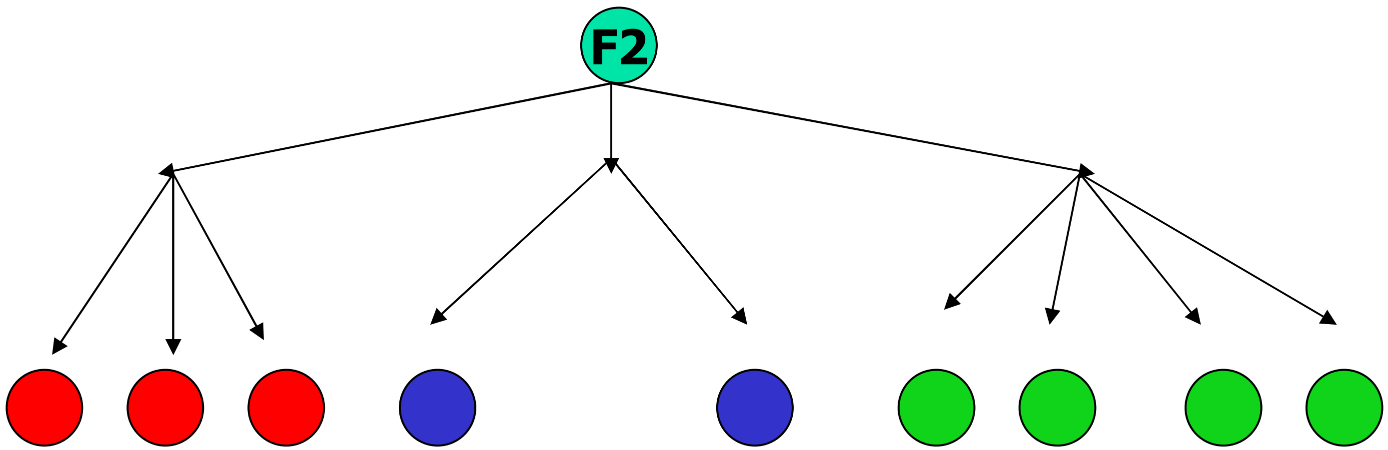
DV method



(any language from group for which documentation available)

DV method

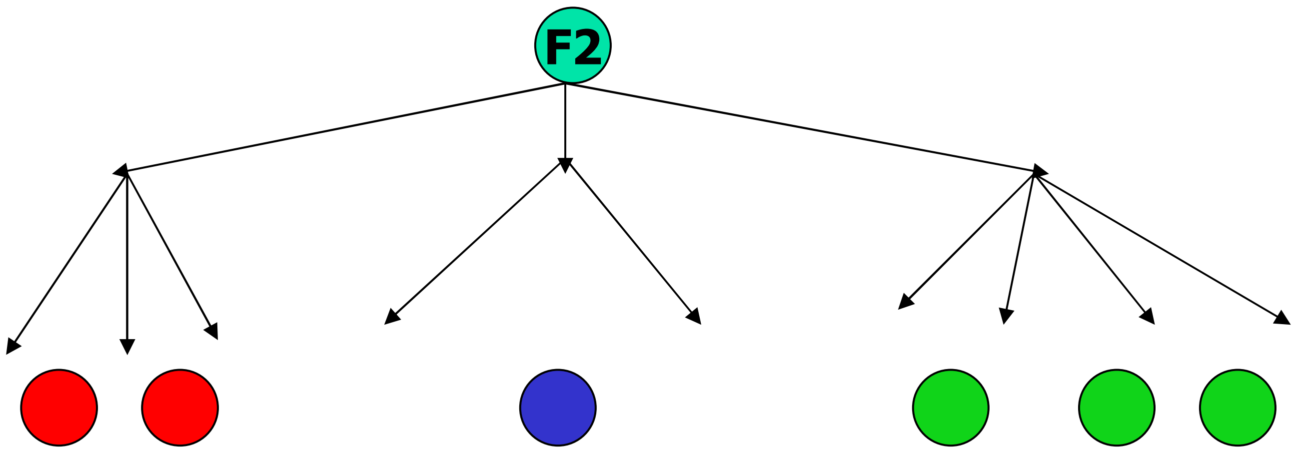
Sample size >>> n



(any language for which documentation available)

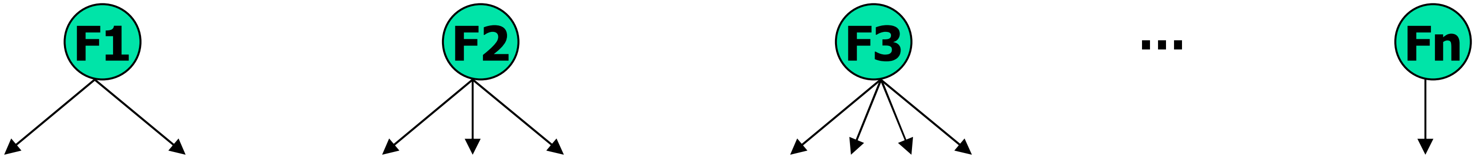
DV method

Sample size >> n



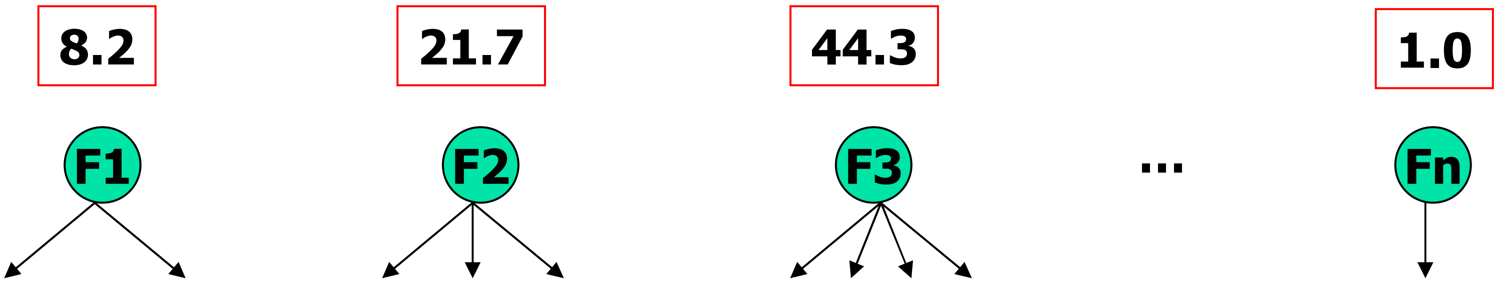
(any language for which documentation available)

DV method



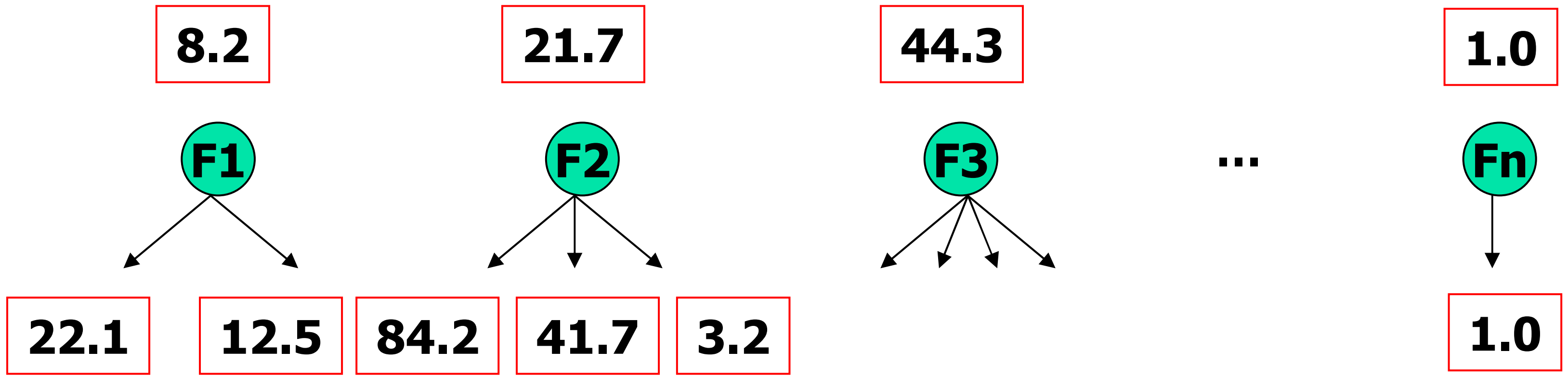
Diversity Value per node, based on:
- **NOT number of daughter nodes**
- **NOT on number of daughter languages, but**
- **Internal Complexity (breadth per level, diminishing)**

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

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

Afro-Asiatic (55.53/6/258)         2
Altaic (15.07/2/62)                1
Korean-Japanese (3.00/3/4)         1
Australian (67.58/30/262)          3
Austric (137.41/3/1186)            5
    Austro-Tai (106.03/2/1027)           3
        Austronesian (118.17/4/970)            2
        Daic (4.67/2/57)                       1
    Austroasiatic (28.08/2/155)          1
    Miao-Yao (2.00/2/4)                  1
…
```

DV method: results

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

Afro-Asiatic (55.53/6/258)          2
Altaic (15.07/2/62)                 1
Korean-Japanese (3.00/3/4)          1
Australian (67.58/30/262)           3
Austric (137.41/3/1186)             5
    Austro-Tai (106.03/2/1027)            3
        Austronesian (118.17/4/970)             2
        Daic (4.67/2/57)                        1
    Austroasiatic (28.08/2/155)           1
    Miao-Yao (2.00/2/4)                   1
…
```

DV method: results

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

Afro-Asiatic (55.53/6/258)        2
Altaic (15.07/2/62)               1
Korean-Japanese (3.00/3/4)        1
Australian (67.58/30/262)         3
Austric (137.41/3/1186)           5
    Austro-Tai (106.03/2/1027)          3
        Austronesian (118.17/4/970)           2
        Daic (4.67/2/57)                      1
    Austroasiatic (28.08/2/155)         1
    Miao-Yao (2.00/2/4)                 1
…
```

DV method

Options:

DV method

Options:

1. Random selection of languages under nodes

DV method

```
Classification: Ethnologue15
Criterion 1: Diversity Value
   Sample size: 150 ( 2.06 % of 7299)

   Austronesian (192.99/12/1268)   5
       Unclassified (1.00/0/1)        1
           1. Ketangalan (G)
       East Formosan (3.00/3/5)       1
           Central (1.00/0/2)             1
           2. Amis          (G)
       Bunun (1.00/0/1)               1
           3. Bunun         (G)
       Western Plains (2.00/2/2)      1
           Thao (1.00/0/1)                1
           4. Thao          (G)
...
```

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:
Bakker, D. (2008).
'LINFER: inferring implications from the WALS database'.
***STUF* 61-3, 186-198.**

UNIVERSALS

Greenberg (1963):

UNIVERSALS

Greenberg (1963):

Absolute: Universal 3

Languages with dominant VSO order are *always* prepositional.

UNIVERSALS

Greenberg (1963):

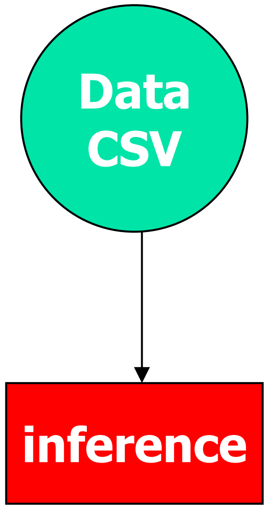
Absolute: Universal 3

Languages with dominant VSO order are *always* prepositional.

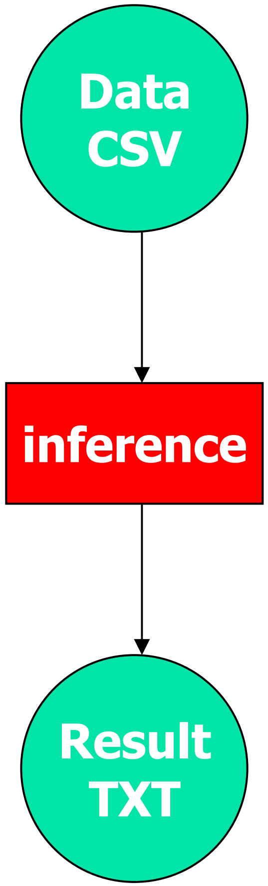
Statistical: Universal 4

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

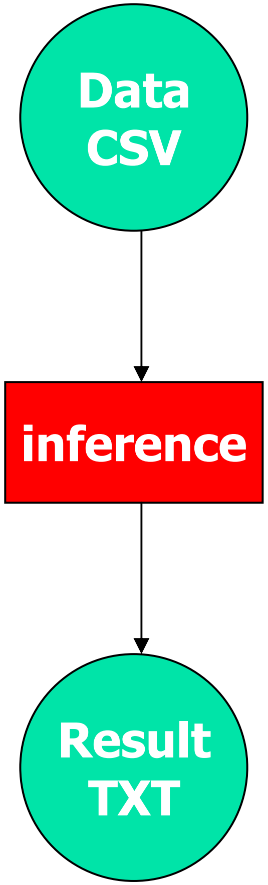
LINFER



LINFER

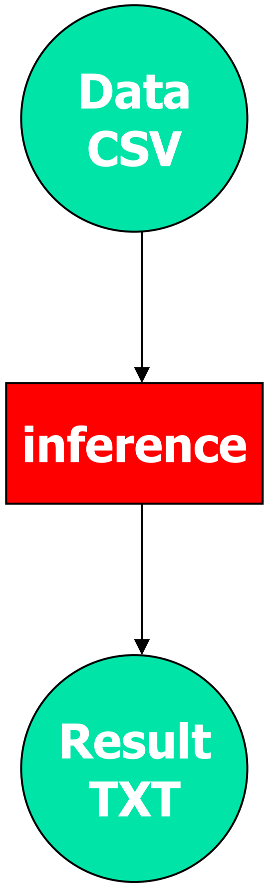


LINFER



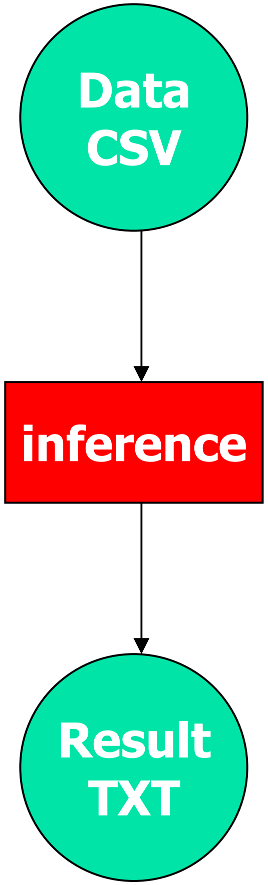
LGNAME				
Abipon				
Abkhaz				
Abun				
Aceh				
Achuma				

LINFER



LGNAME	SmrkP	SmrkV	SmrkN	SmrkH
Abipon	123	No	Sgpl	No
Abkhaz	123	No	Sgpl	No
Abun	12	No	Sg	No
Aceh	12	Yes	Nonum	Irr
Achuma	123	No	Sgdupl	Yes

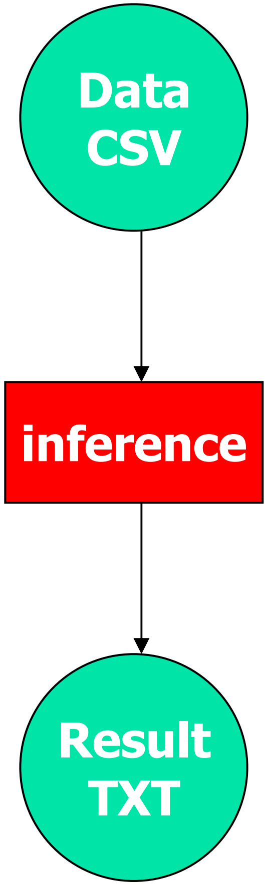
LINFER



LGNAME	**SmrkP** →	**SmrkV**	SmrkN	SmrkH
Abipon	**123**	**No**	Sgpl	No
Abkhaz	**123**	**No**	Sgpl	No
Abun	12	No	Sg	No
Aceh	12	Yes	Nonum	Irr
Achuma	**123**	**No**	Sgdupl	Yes

SmrkP = 123 → SmrkV = No (ABS)

LINFER



LGNAME	SmrkP ←	SmrkV	SmrkN	SmrkH
Abipon	123	No	Sgpl	No
Abkhaz	123	No	Sgpl	No
Abun	12	No	Sg	No
Aceh	12	Yes	Nonum	Irr
Achuma	123	No	Sgdupl	Yes

SmrkP = 123 → SmrkV = No (ABS)

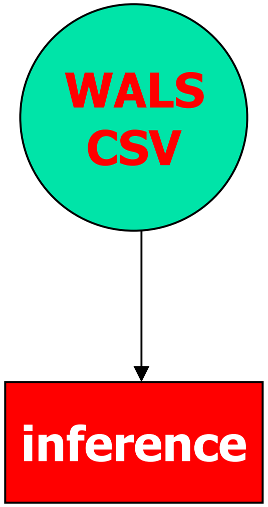
SmrkV = No → SmrkP = 123 (0.75)

LINFER

- Automatic inference of implications (A & S):

→ **generate + test**

LINFER: on WALS



Haspelmath, M., M. Dryer, D. Gil & B. Comrie (eds) (2005).
The World Atlas Of Language Structures.
Oxford: Oxford University Press
WALS Online: http://wals.info/

Number of languages: 2558
Number of variables: 143

LINFER

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%)

LINFER

1. CORSEX=1 <=> CORNUM=1 n=144
   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

LINFER

TRIVIAL!!!

1. CORSEX=1 CORNUM=1 n=144

0.56 [3

[Fr=0.563, Fa=1.00 Fn=1.000, chi2<0.5%] EQUIV

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

<=>

Number of Genders: None

LINFER

15. KAYBCC=4 <=> VESTAM=2 n= 2
0.13 [7] - 0.13 [4]

[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
<=>
Suppletion According to Tense and Aspect:
Suppletion according to aspect

LINFER

15. KAYBCC= … ESTAM=2   n= 2
    0.13 […]

[Fr=0.133, Fa=1.000, … n=1.000, chi2<0.5%] EQUIV

INSIGNIFICANT!!!

Number of Basic Colour Cate…
  7 or between 7 and 8 categori…
      <=>
Suppletion According to Tense and Aspect
  Suppletion according to aspect

LINFER

57. DRYRPO=4 => DRYREL=1 n=291
    0.49 [5] - 0.73 [7]

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

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

LINFER

57. DRYRPO=4 => DRYREL=1 n=291
    0.49 [5] – 0.73 [7]

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

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

VO & Prep → NRel

LINFER

57. DRYRPO=4 => DRYREL=1 n=291
0.49 [5] - 0.73 [7]

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

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

VO & Prep → NRel

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

LINFER

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

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

EXC: fre grb lkt

LINFER

```
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 → NRel

LINFER

<u>Two major questions:</u>

LINFER

Two major questions:

1. When is an implication statistically reliable?

LINFER

Two major questions:

1. When is an implication statistically reliable?

2. When is an implication linguistically interesting?

LINFER

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 / Σp_i)
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 >

LINFER

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 ($p / \Sigma p_i$)
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 x m tables (not tetrachoric)
Fisher Exact: when tetrachoric and 1 empty cell
Other statistics: <export data >

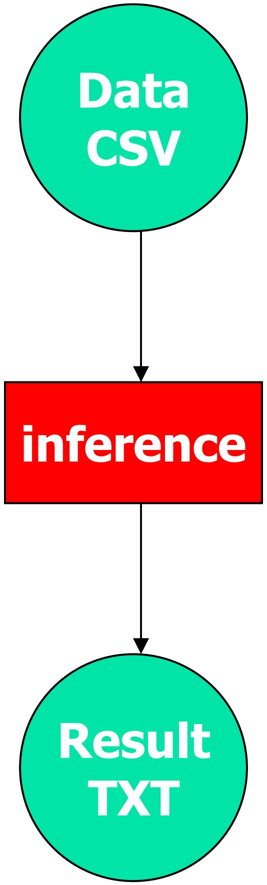
LINFER

57. DRYRPO=4 => DRYREL=1   n=291

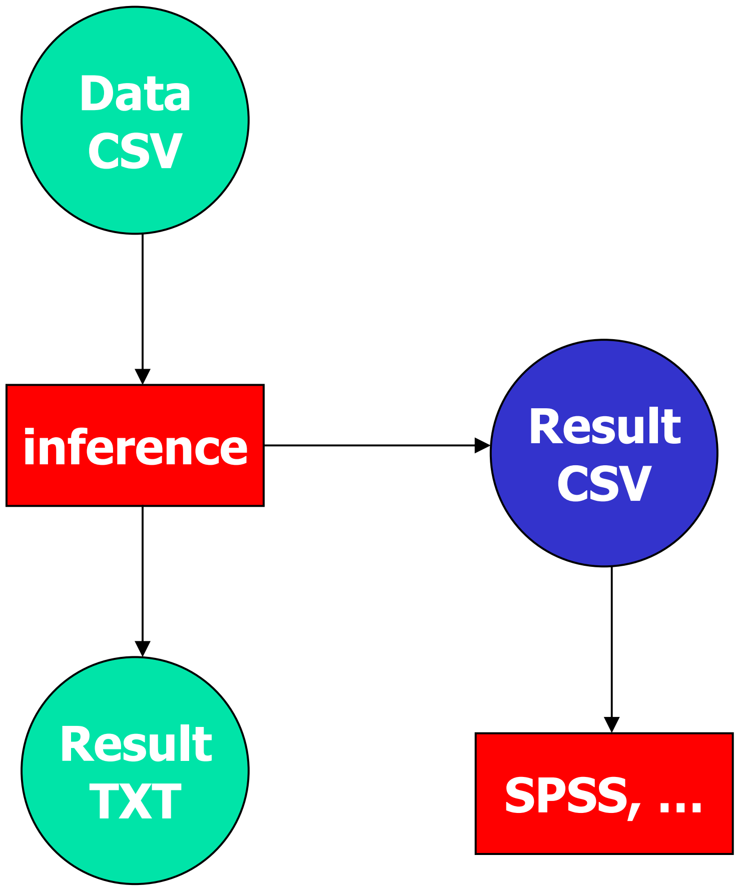
[Fr=0.486, Fa=0.983, Fc=0.655, Fn=0.511, chi2<0.5%] STAT

Relevance: proportion of values for premisse (p / Σp_i)
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 >

LINFER



LINFER



LINFER: DIACHRONY?

DIACHRONY?:

```
99. CYSVRB=3 => CYSIND=3     n= 75
      0.40 [5] - 0.60 [5]

[Fr=0.395, Fa=0.949, Fc=0.625, Fn=0.628, chi2<0.5%] STAT

Inclusive/Exclusive Distinction in Verba:
  No inclusive/exclusive opposition
        =>
Inclusive/Exclusive Distinction in Pronoun:
  No inclusive/exclusive opposition

EXC: abk cle map mrd
```

LINFER: RELATED?

RELATED?:

```
187. HASNPL=6 => BROFIN=2     n= 49
     0.49 [6] - 0.93 [2]

[Fr=0.490, Fa=1.000, Fc=0.527, Fn=0.137, chi2<1.0%] ABS

Occurrence of Nominal Plurality:
  Plural in all nouns, always obligatory
        =>
Finger and Hand:
  Different words denote 'hand' and `finger'
```

LINFER

WALS(-like) database, observations:

LINFER

WALS(-like) database:

- Less than 1:1000 logically possible implications are of potential interest

LINFER

WALS(-like) database:

- Less than 1:1000 logically possible implications are of potential interest
- Most equivalences are trivial

LINFER

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

LINFER

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

ASJP

Project ASJP (= Automated Similarity Judgment Program)

ASJP

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

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

<u>Reference:</u>
Bakker, D., A. Müller, V. Velupillai, S. Wichmann, C. H. Brown, P. Brown, D. Egorov, R. Mailhammer, A. Grant, E. W. Holman (2009). 'Adding typology to lexicostatistics: a combined approach to language classification'. *Linguistic Typology* 13, 167-179.

ASJP

Project:
ASJP (Automated Similarity Judgment Program)

Overall goal:
Automatic reconstruction of language relationships
(lexical, grammatical → genetic, areal, typological, ...)

ASJP

Project:
ASJP (Automated Similarity Judgment Program)

Overall goal:
Automatic reconstruction of language relationships

Basis:
Distance matrix between individual languages based on lexical features

ASJP

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,

ASJP

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

ASJP

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

ASJP

Project:
ASJP (Automated Similarity Judgment Program)

As in traditional lexicostatistics, but:

1. use of computational algorithms and tools

2. methodology from classification in biology

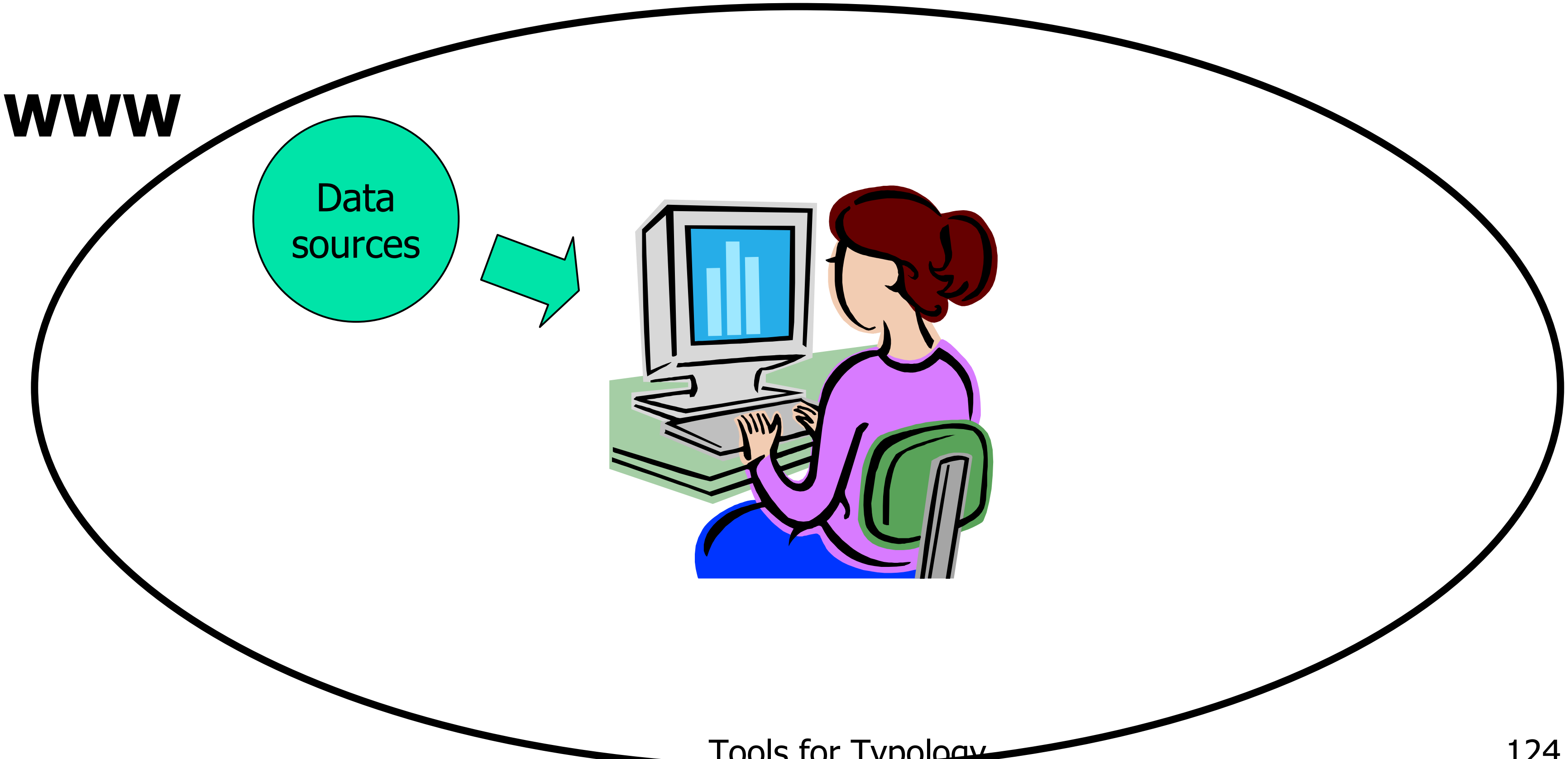
ASJP

Project:
ASJP (Automated Similarity Judgment Program)

WWW

ASJP

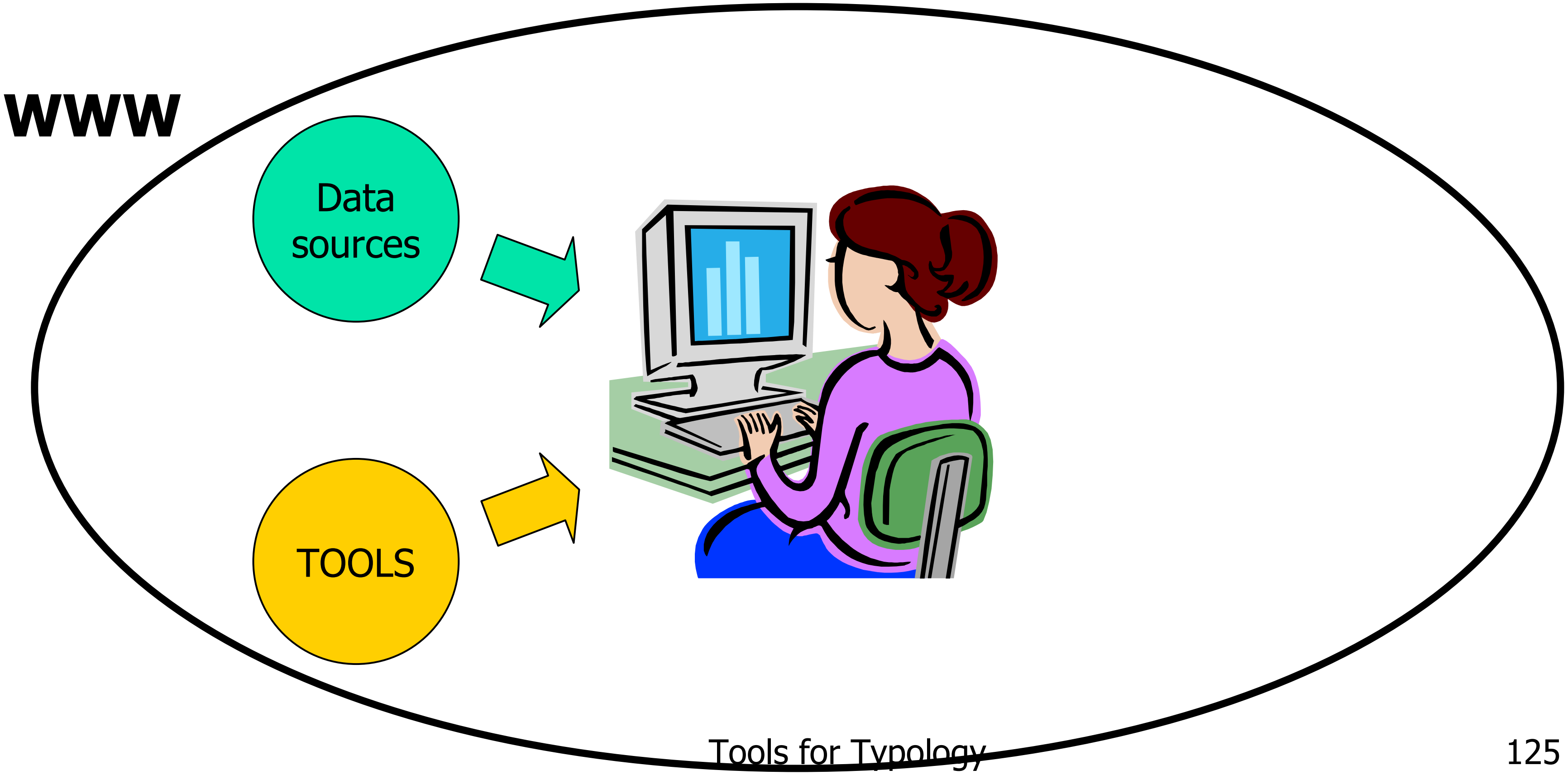
Project:
ASJP (Automated Similarity Judgment Program)



ASJP

Project:

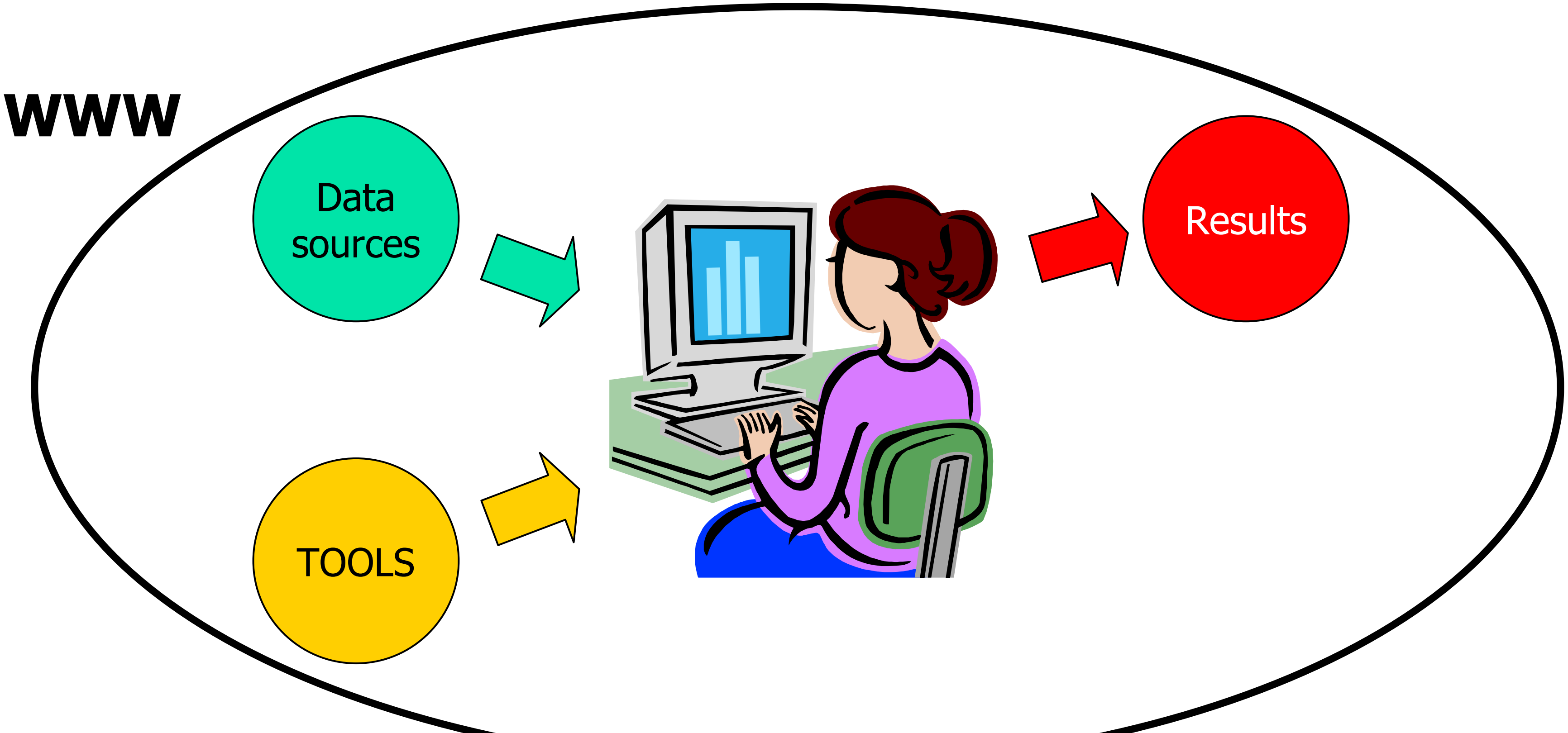
ASJP (Automated Similarity Judgment Program)



ASJP

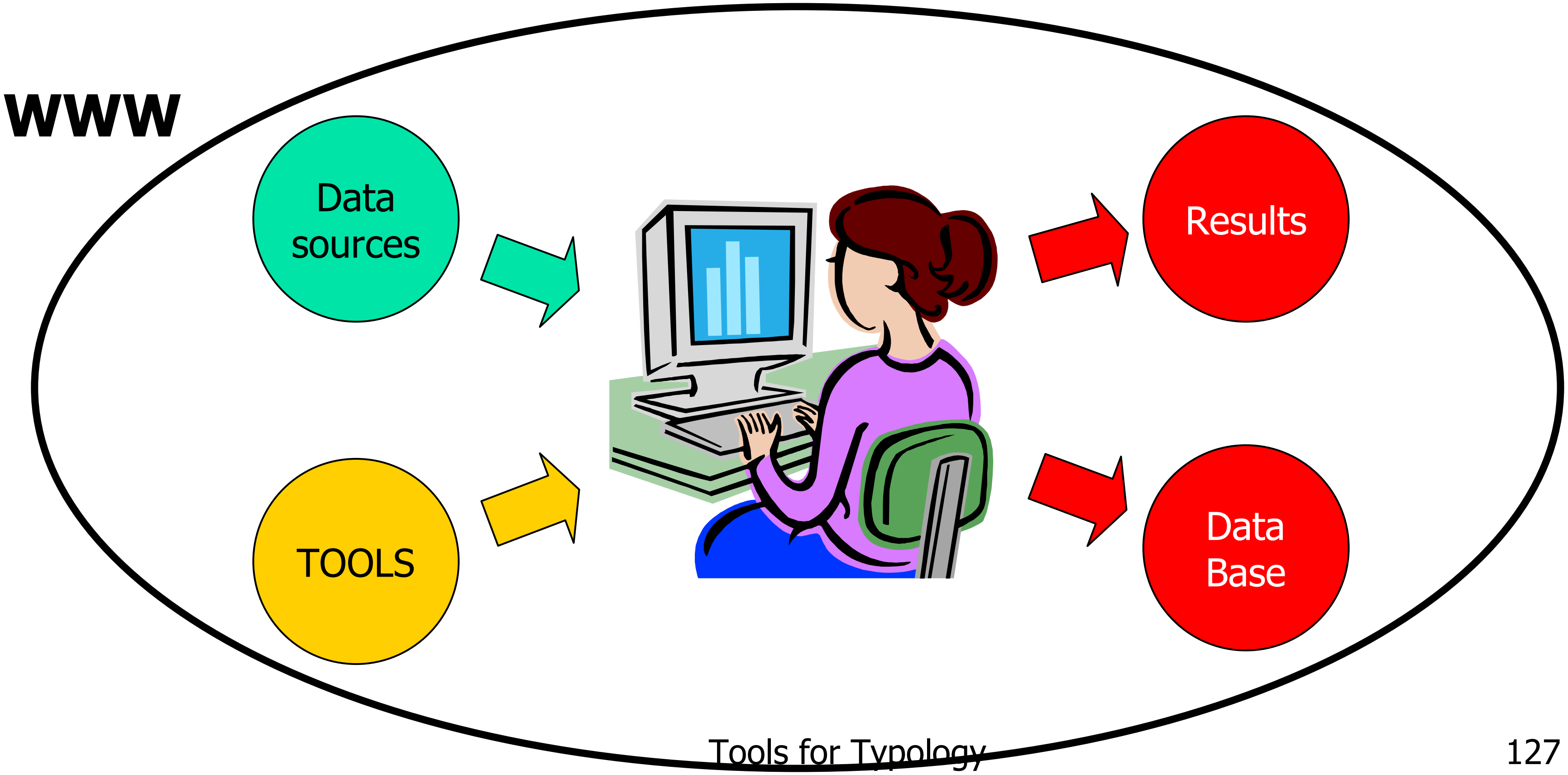
Project:

ASJP (Automated Similarity Judgment Program)



ASJP

Project:
ASJP (Automated Similarity Judgment Program)

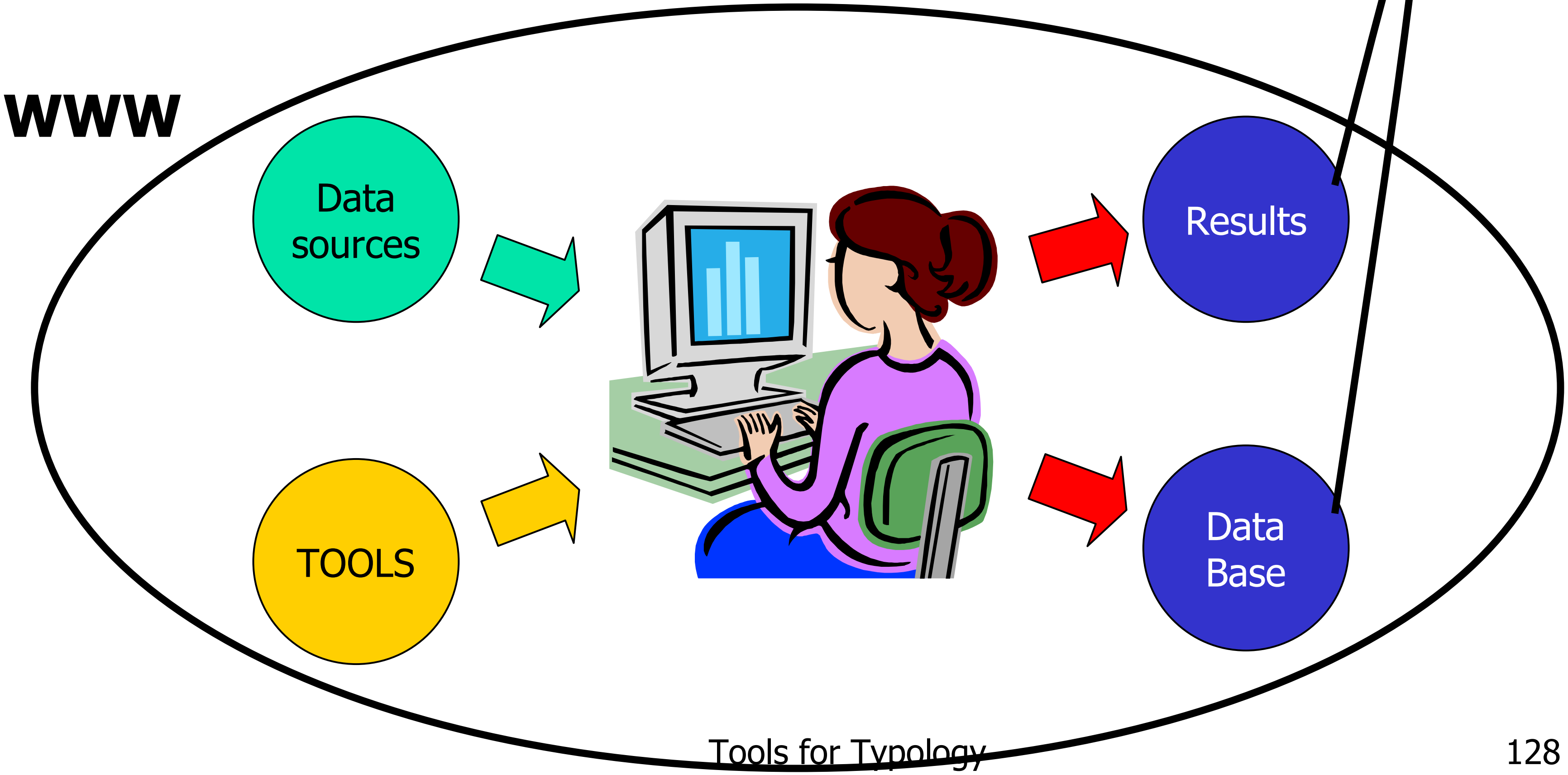


ASJP

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

Project:

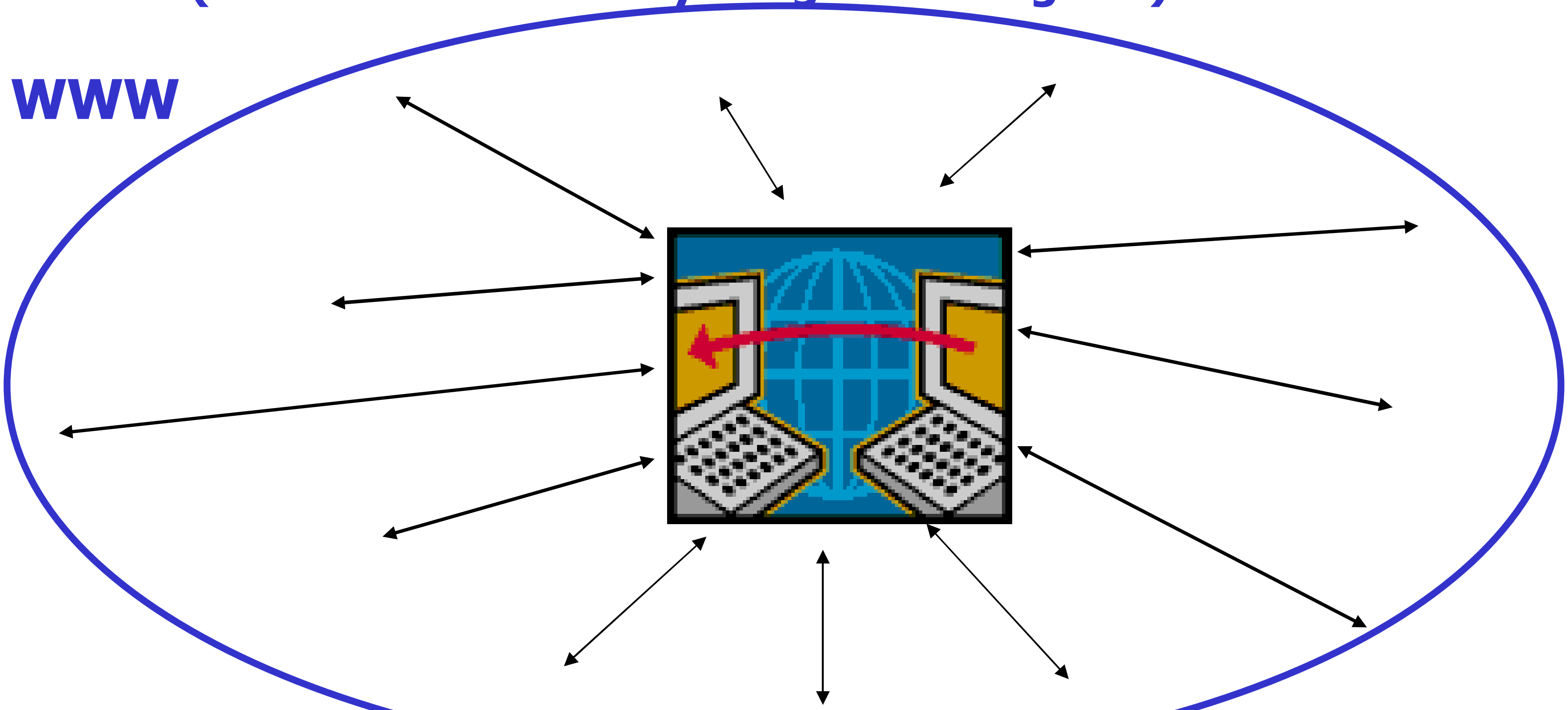
ASJP (Automated Similarity Judgment Program)



ASJP

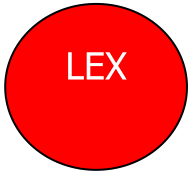
Project:
ASJP (Automated Similarity Judgment Program)

WWW

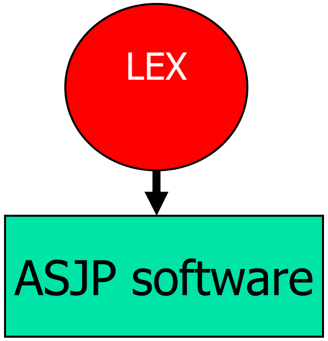


Overview ASJP system

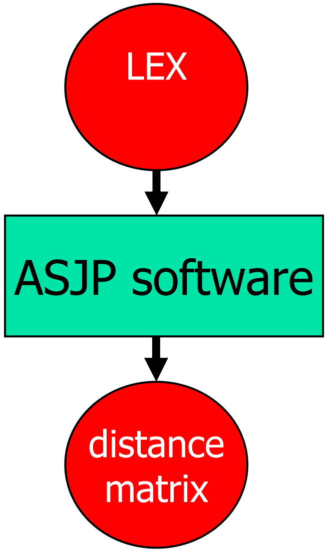
Overview ASJP system



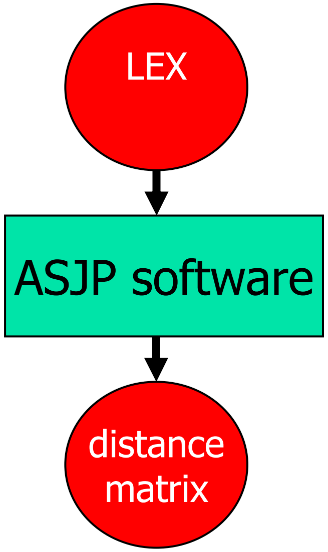
Overview ASJP system



Overview ASJP system

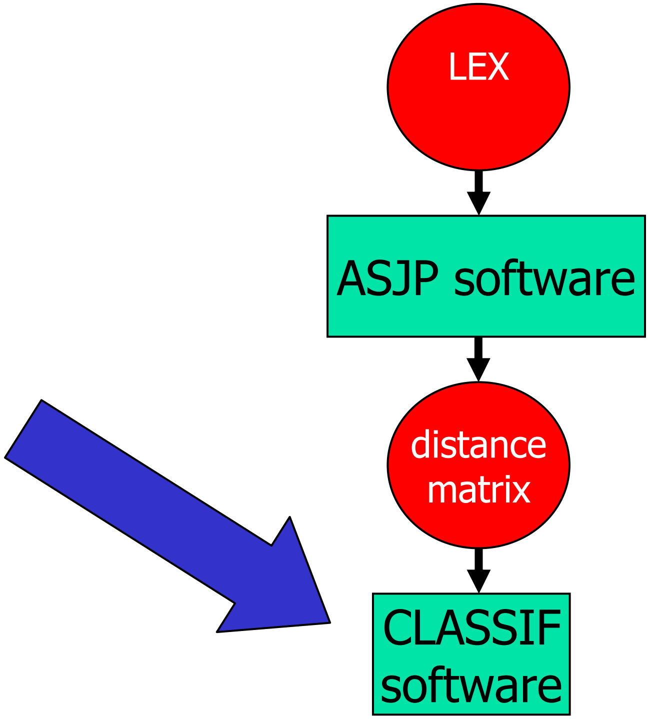


Overview ASJP system

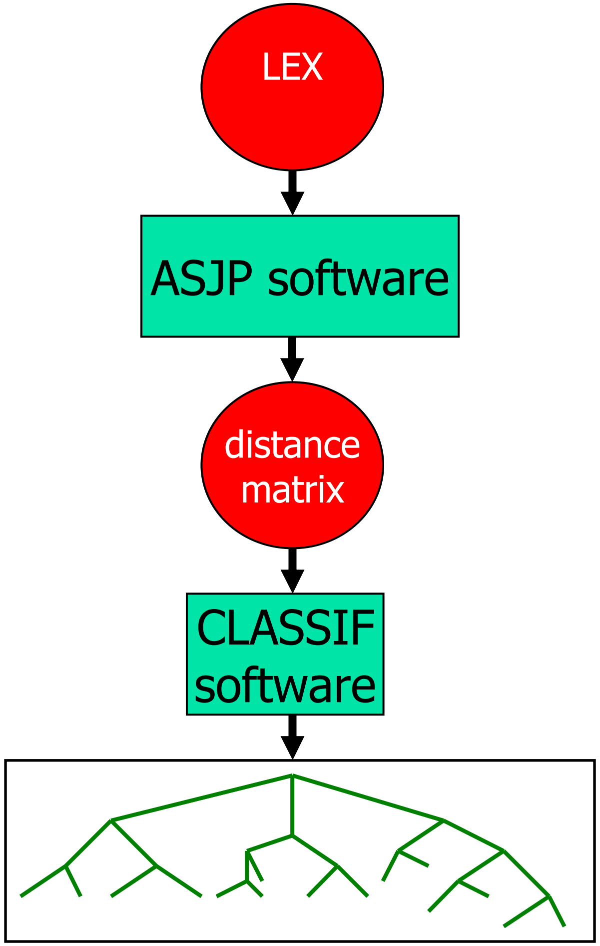


DUTCH	ENGLISH	53.3
DUTCH	FRENCH	72.7
DUTCH	MANDARIN	93.8
…		

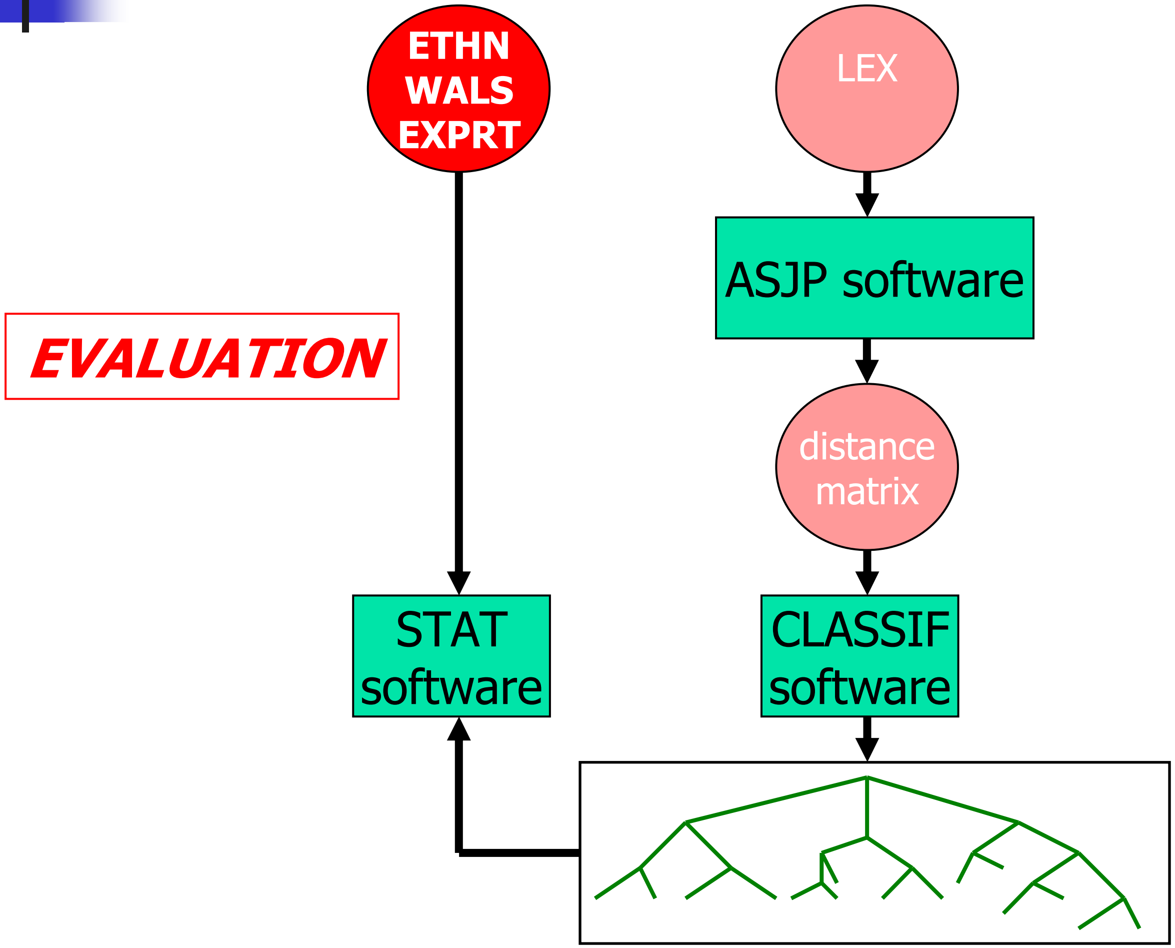
Overview ASJP system



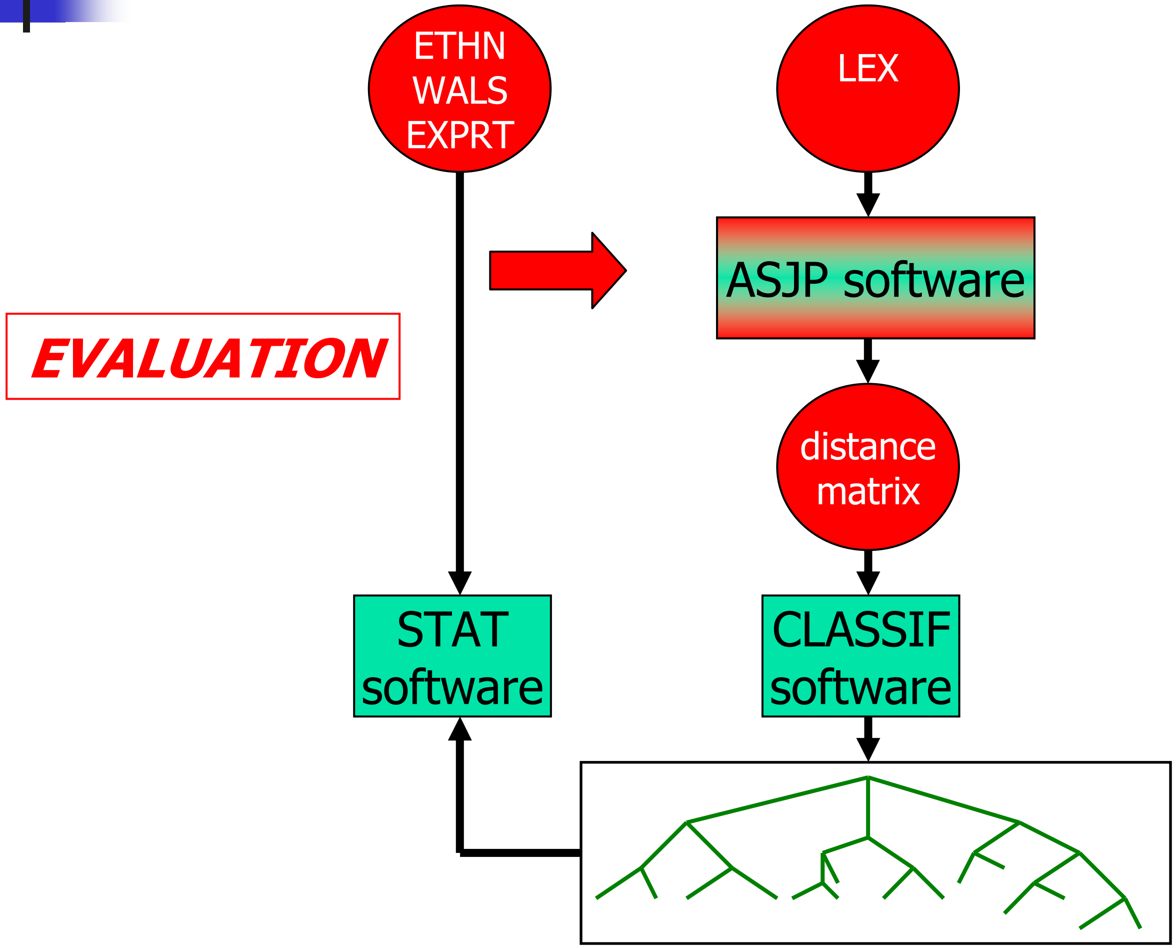
Overview ASJP system

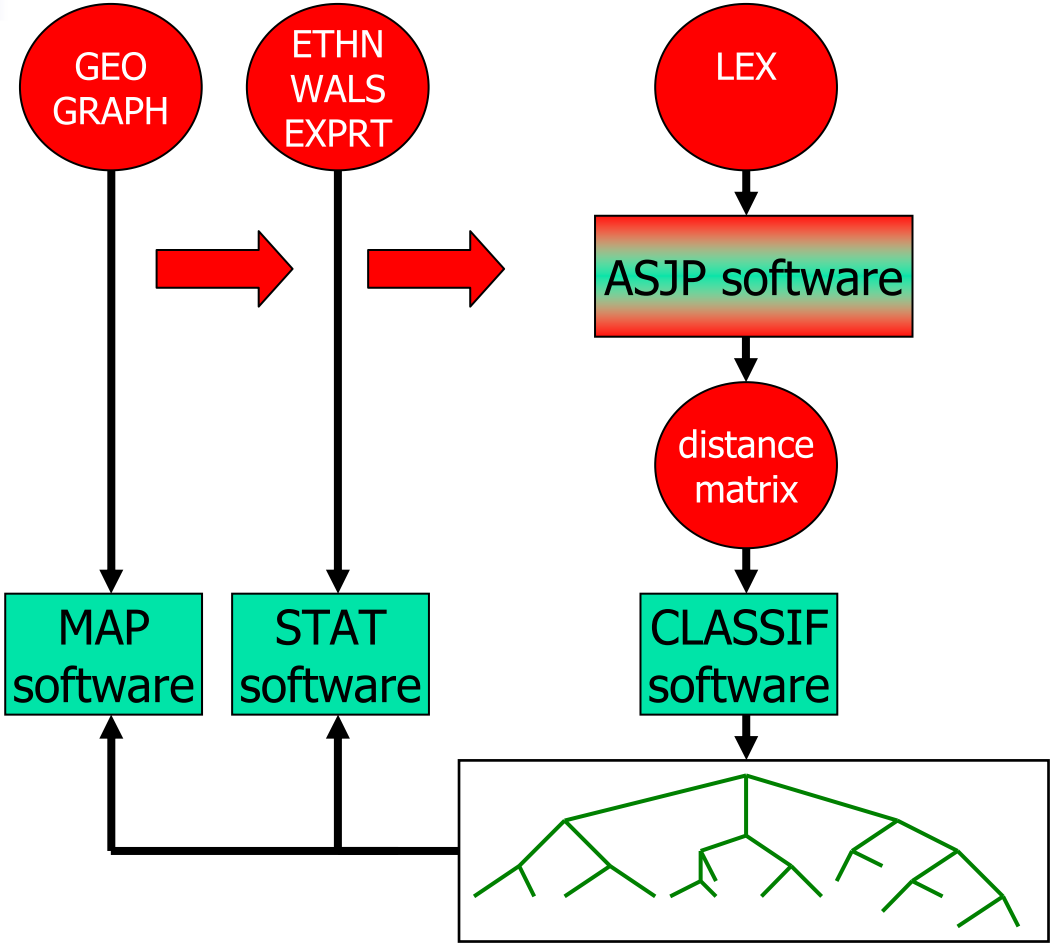


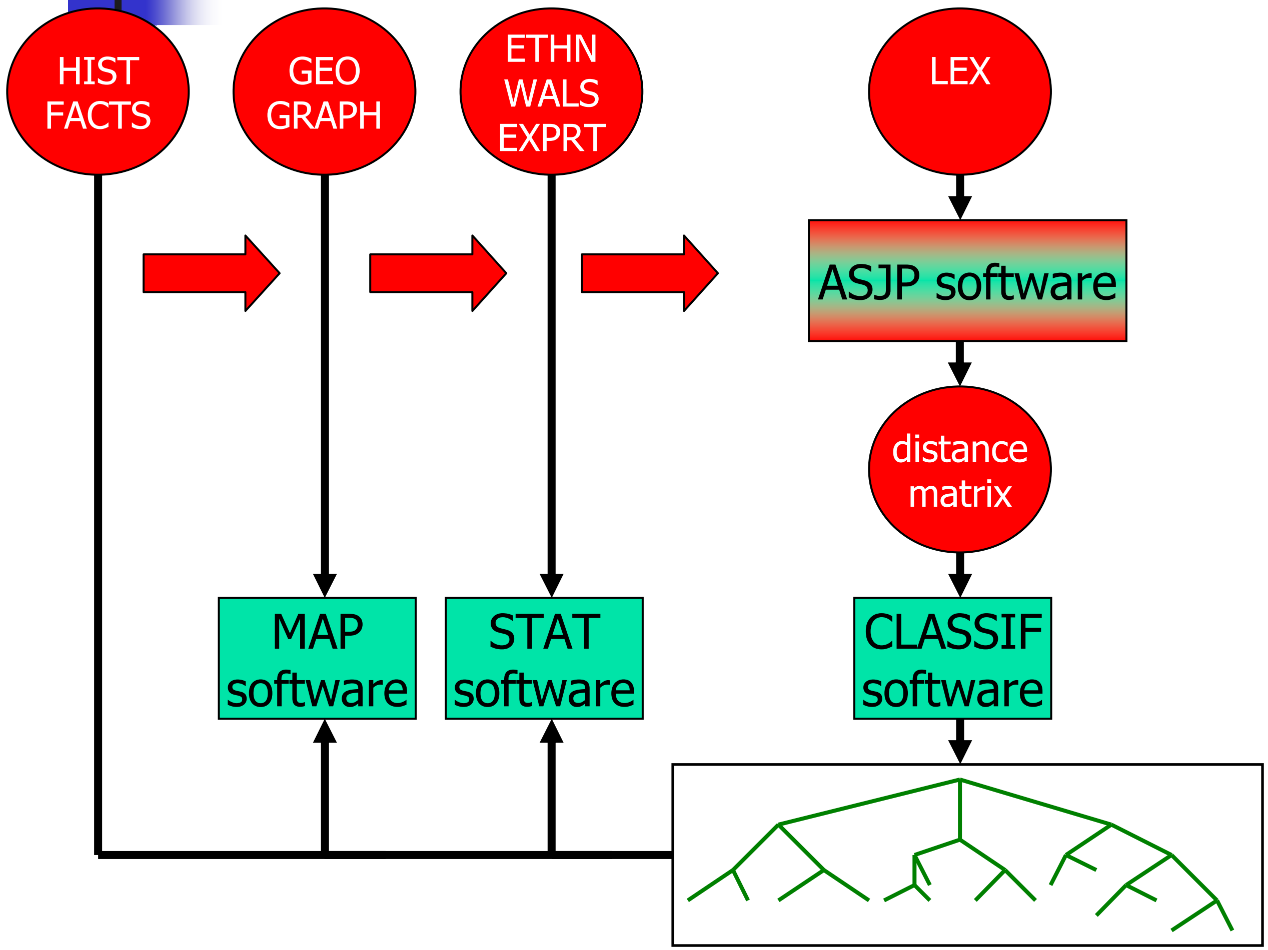
Existing Expert Classifications:

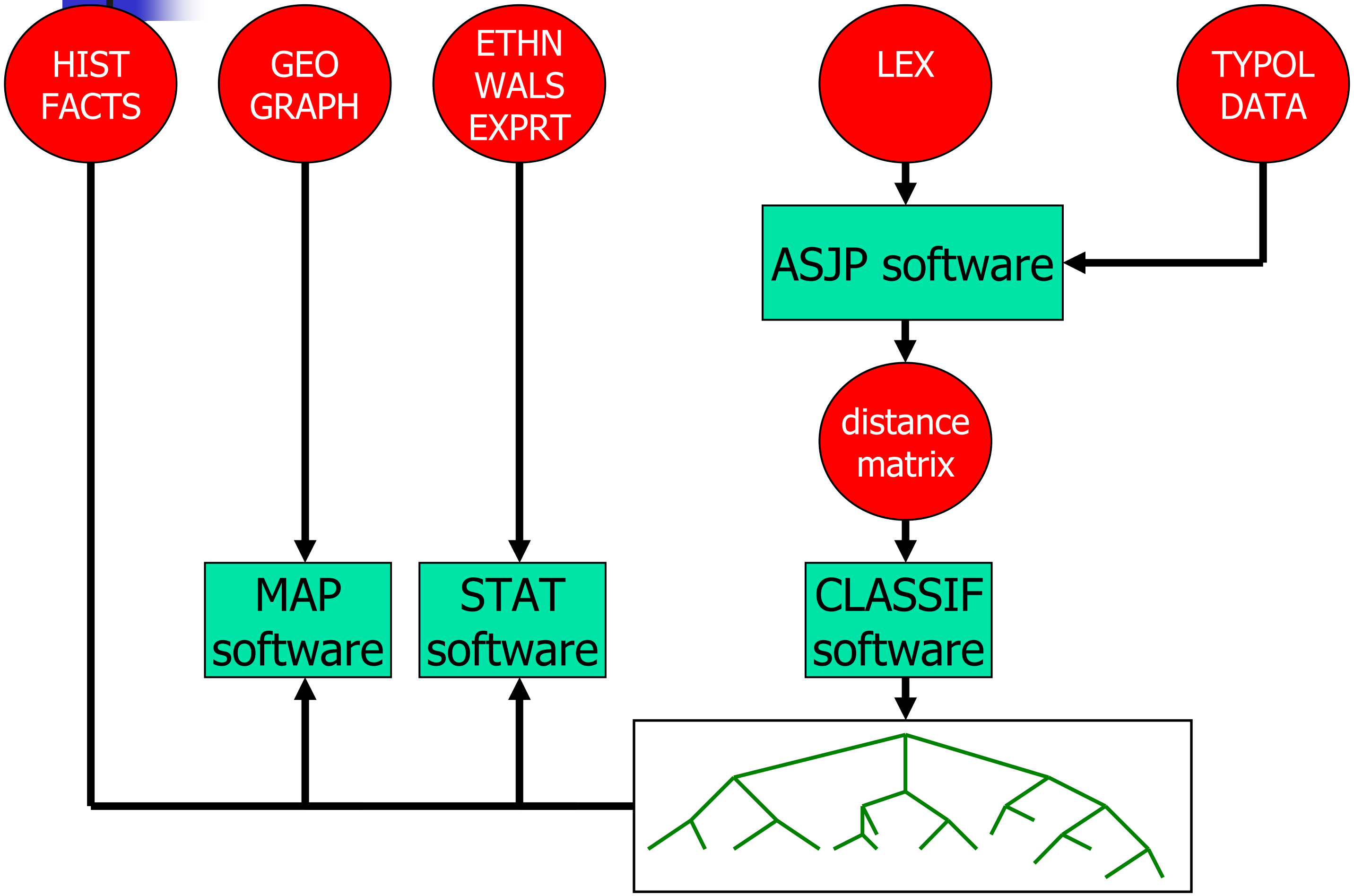


Existing Expert Classifications:




GEO GRAPH
ETHN WALS EXPRT
LEX
ASJP software
distance matrix
MAP software
STAT software
CLASSIF software


HIST FACTS
GEO GRAPH
ETHN WALS EXPRT
LEX
ASJP software
distance matrix
MAP software
STAT software
CLASSIF software


HIST
FACTS
GEO
GRAPH
ETHN
WALS
EXPRT
LEX
TYPOL
DATA
ASJP software
distance
matrix
MAP
software
STAT
software
CLASSIF
software

Lexical items

Data: Word list Morris Swadesh (1955):

100 basic meanings

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

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)

GERMANIC	FISH
AFRIKAANS	fis
BERNESE_GERMAN	fiS
BRABANTIC	fis
CIMBRIAN	fiS
DANISH	fesk
DUTCH	vis
ENGLISH	fiS
FAROESE	fiskur
FRANS_VLAAMS	fiS
FRISIAN_WESTERN	fisk
GOTHIC	fisks
ICELANDIC	fiskir
JAMTLANDIC	fisk
LIMBURGISH	vES
LUXEMBOURGISH	feS
NORTH_FRISIAN_AMRUM	fask

GERMANIC	FISH
AFRIKAANS	fis
BERNESE_GERMAN	fiS
BRABANTIC	fis
CIMBRIAN	fiS
DANISH	fesk
DUTCH	vis
ENGLISH	fiS
FAROESE	fiskur
FRANS_VLAAMS	fiS
FRISIAN_WESTERN	fisk
GOTHIC	fisks
ICELANDIC	fiskir
JAMTLANDIC	fisk
LIMBURGISH	vES
LUXEMBOURGISH	feS
NORTH_FRISIAN_AMRUM	fask

1 proto form

GERMANIC	TREE
AFRIKAANS	bom
BERNESE_GERMAN	boum
BRABANTIC	bu3m
DANISH	trE7
DUTCH	bom
ENGLISH	tri
FAROESE	trEa
FRANS_VLAAMS	bom
FRISIAN_WESTERN	bi3m|by~Em
GOTHIC	bagms|triu
ICELANDIC	th~ry~E
JAMTLANDIC	tre
LIMBURGISH	boum
LUXEMBOURGISH	bam
NORTH_FRISIAN_AMRUM	bum
NORTHERN_LOW_SAXON	bom
NORWEGIAN_BOKMAAL	tre

GERMANIC	TREE
AFRIKAANS	bom
BERNESE_GERMAN	boum
BRABANTIC	bu3m
DANISH	trE7
DUTCH	bom
ENGLISH	tri
FAROESE	trEa
FRANS_VLAAMS	bom
FRISIAN_WESTERN	bi3m|by~Em
GOTHIC	bagms|triu
ICELANDIC	th~ry~E
JAMTLANDIC	tre
LIMBURGISH	boum
LUXEMBOURGISH	bam
NORTH_FRISIAN_AMRUM	bum
NORTHERN_LOW_SAXON	bom
NORWEGIAN_BOKMAAL	tre

2 forms

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	q3l3
CSANGO	hol
HUNGARIAN	hal

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	q3l3
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

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

4 forms

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:

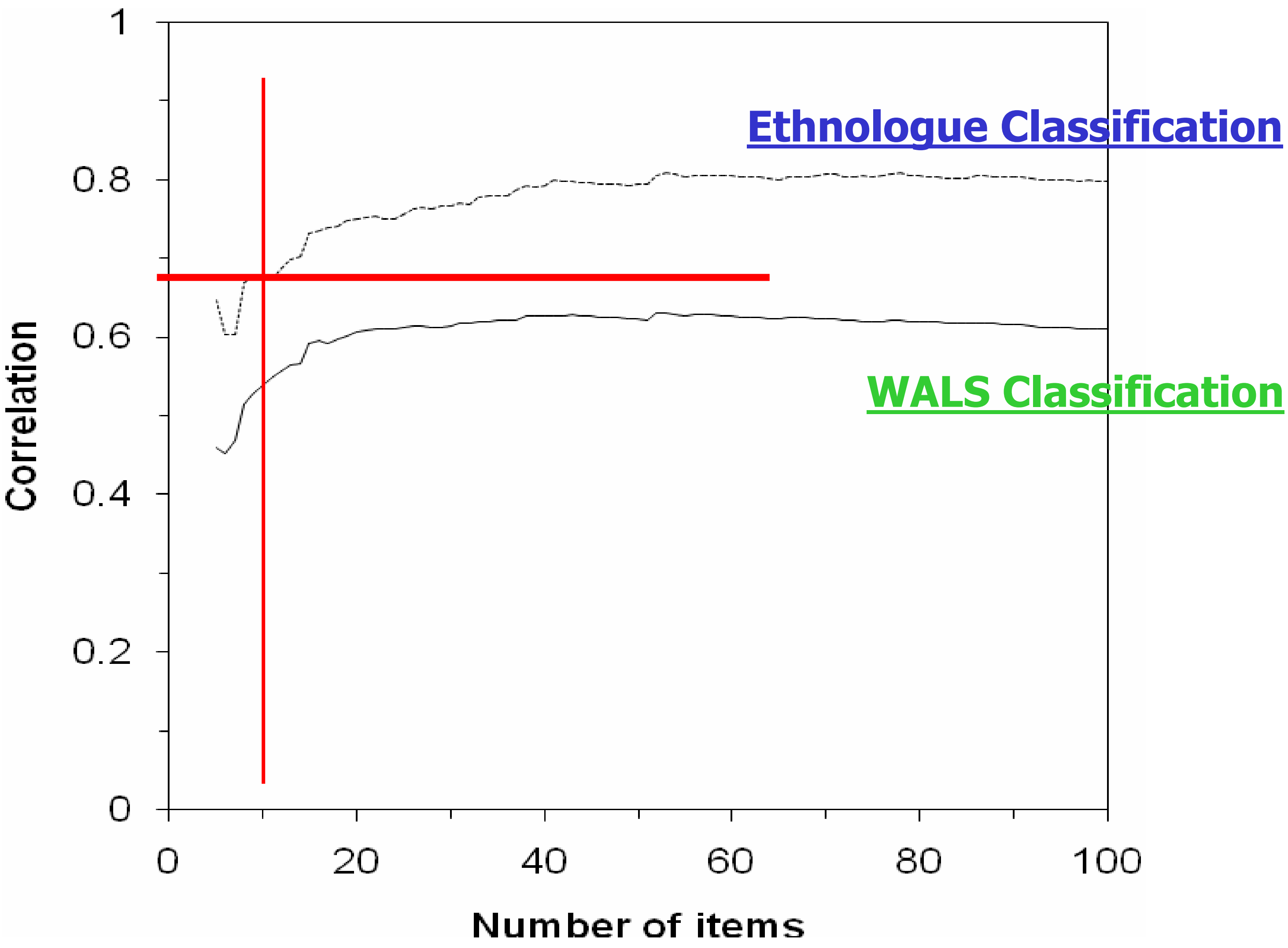
- **at least the same results as > 40**

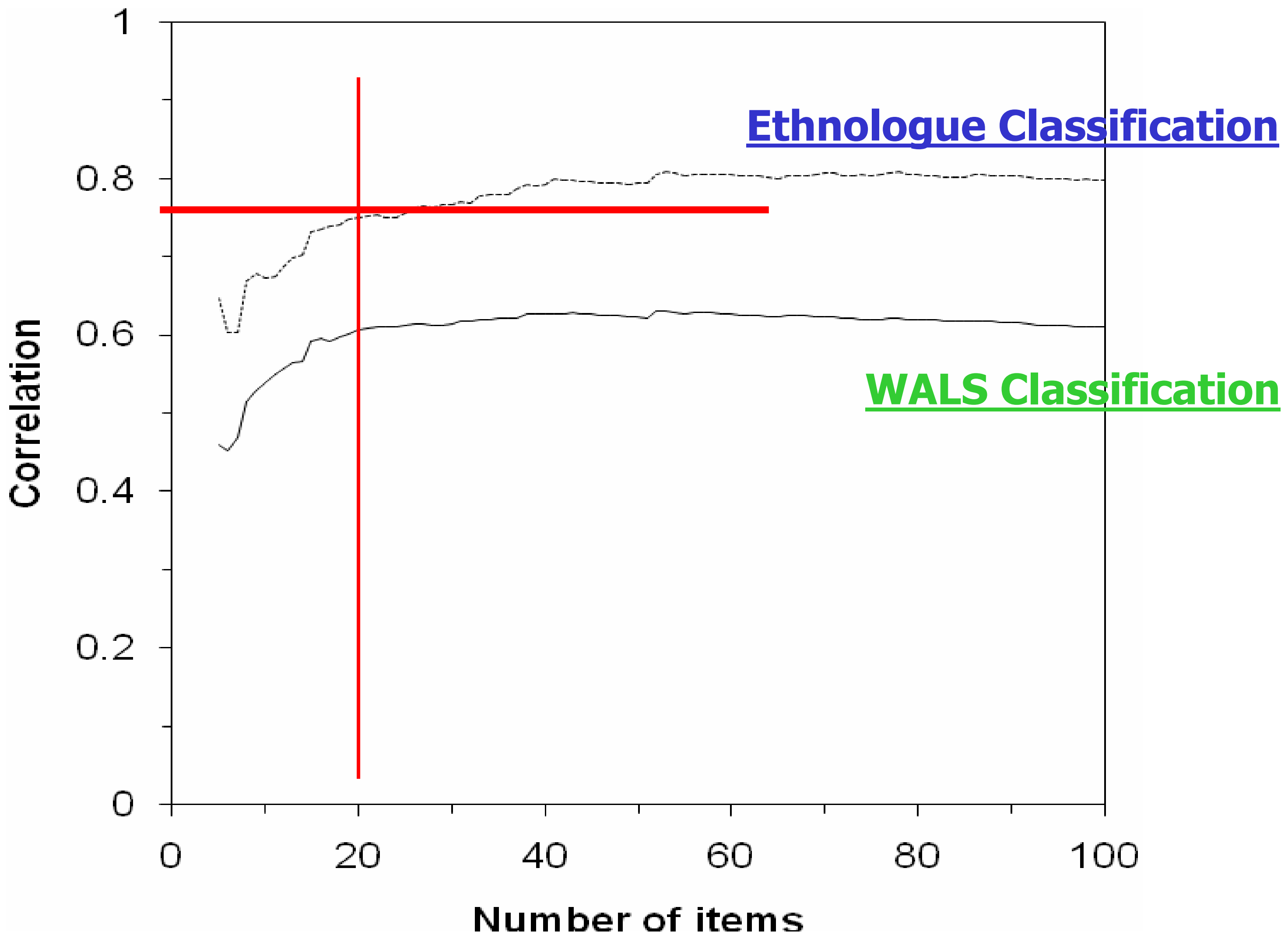
Lexical items: further reduction

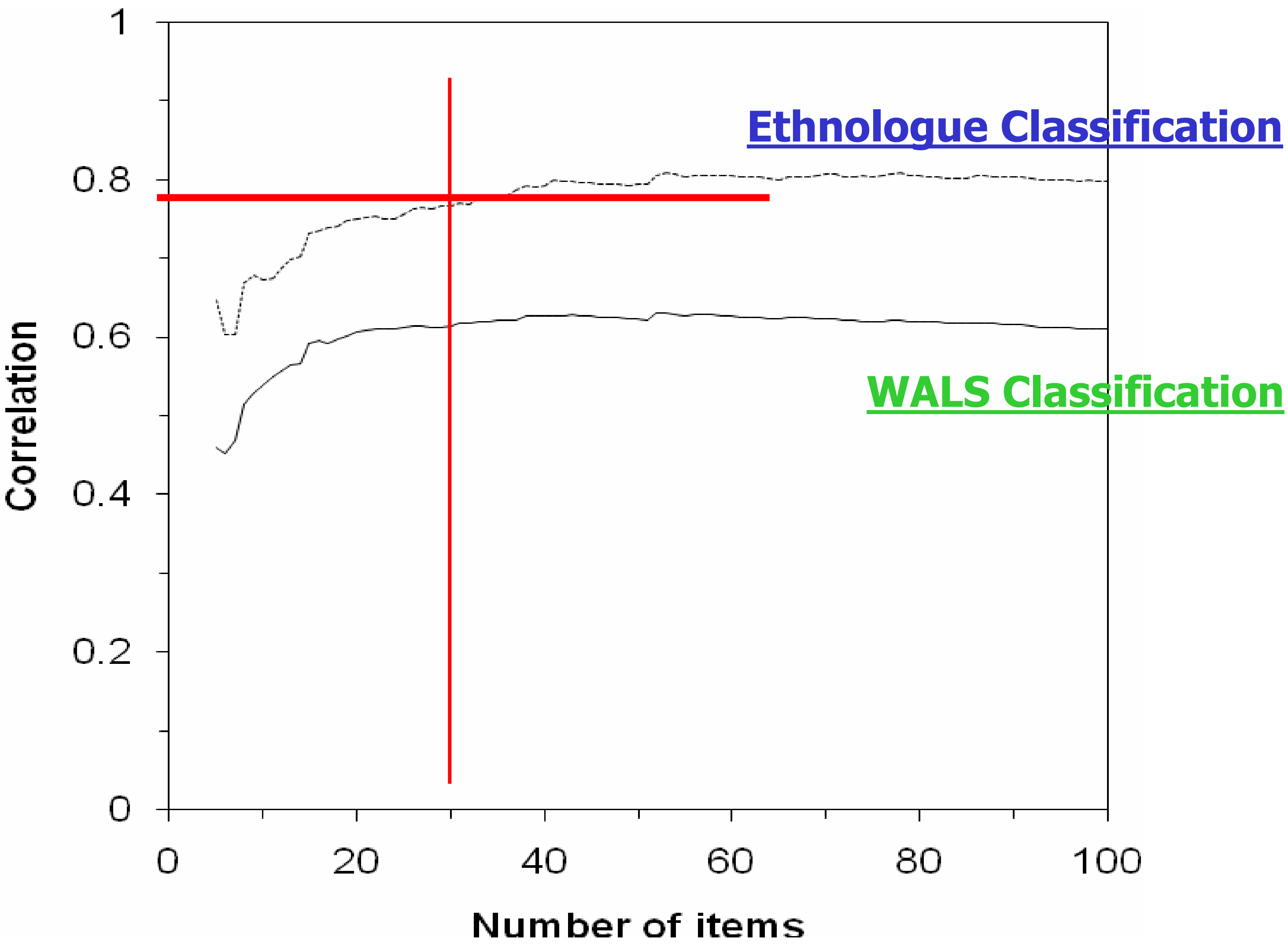
Early analyses have shown:

Most stable 40/100 item subset gives:

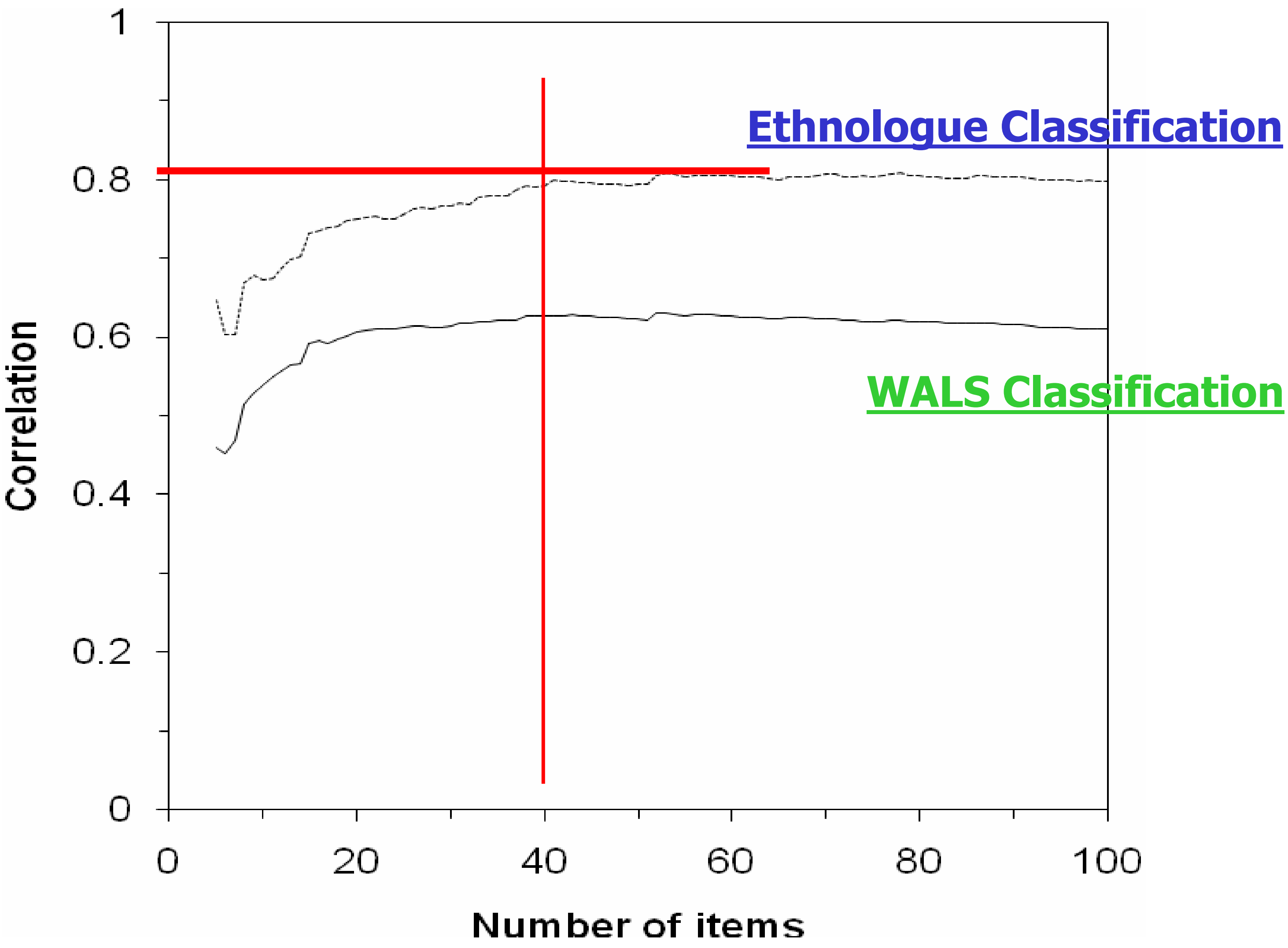
- **at least the same results as > 40**
- **better results than < 40**

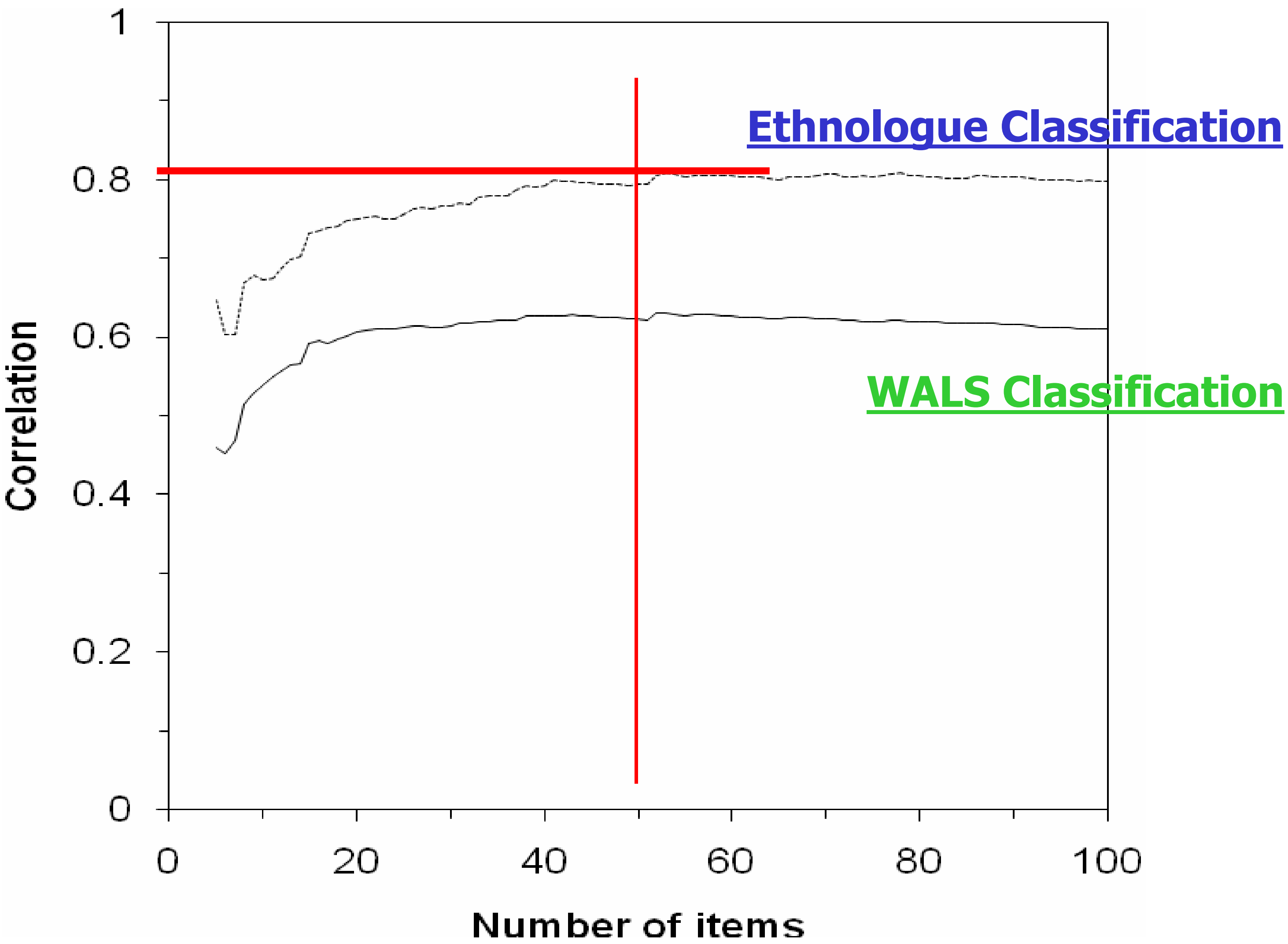

Ethnologue Classification
WALS Classification
1
0.8
0.6
0.4
0.2
0
Correlation
0
20
40
60
80
100
Number of items


Ethnologue Classification
WALS Classification
Correlation
Number of items
1
0.8
0.6
0.4
0.2
0
0
20
40
60
80
100



Ethnologue Classification
WALS Classification
Correlation
Number of items
1
0.8
0.6
0.4
0.2
0
0
20
40
60
80
100


Ethnologue Classification
WALS Classification
Correlation
1
0.8
0.6
0.4
0.2
0
0
20
40
60
80
100
Number of items


1
0.8
0.6
0.4
0.2
0
Correlation
0
20
40
60
80
100
Number of items
Ethnologue Classification
WALS Classification

I	dog	nose	die	smoke
you	louse	mouth	kill	fire
we	tree	tooth	swim	ash
this	seed	tongue	fly	burn
that	leaf	claw	walk	path
who	root	foot	come	mountain
what	bark	knee	lie	red
not	skin	hand	sit	green
all	flesh	belly	stand	yellow
many	blood	neck	give	white
one	bone	breast	say	black
two	grease	heart	sun	night
big	egg	liver	moon	hot
long	horn	drink	star	cold
small	tail	eat	water	full
woman	feather	bite	rain	new
man	hair	see	stone	good
person	head	hear	sand	round
fish	ear	know	earth	dry
bird	eye	sleep	cloud	name

I	dog	nose	die	smoke
you	louse	mouth	kill	fire
we	tree	tooth	swim	ash
this	seed	tongue	fly	burn
that	leaf	claw	walk	path
who	root	foot	come	mountain
what	bark	knee	lie	red
not	skin	hand	sit	green
all	flesh	belly	stand	yellow
many	blood	neck	give	white
one	bone	breast	say	black
two	grease	heart	sun	night
big	egg	liver	moon	hot
long	horn	drink	star	cold
small	tail	eat	water	full
woman	feather	bite	rain	new
man	hair	see	stone	good
person	head	hear	sand	round
fish	ear	know	earth	dry
bird	eye	sleep	cloud	name

40 Most Stable

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

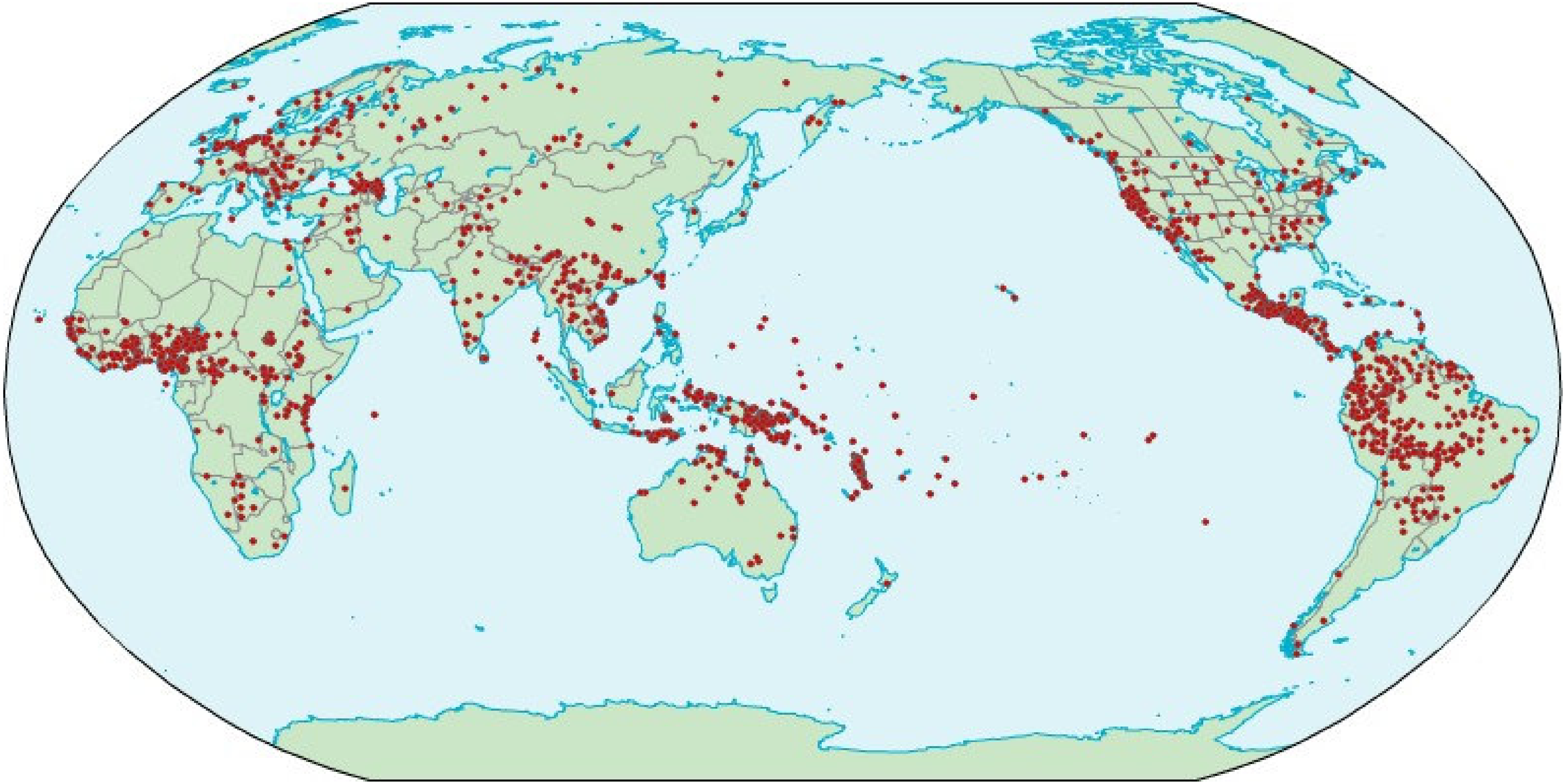
→ Less missing data

→ Faster processing; combinatorial explosion:

$$\boxed{40 : 100 \sim 2.5 * 2.5 = 6.3}$$

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:

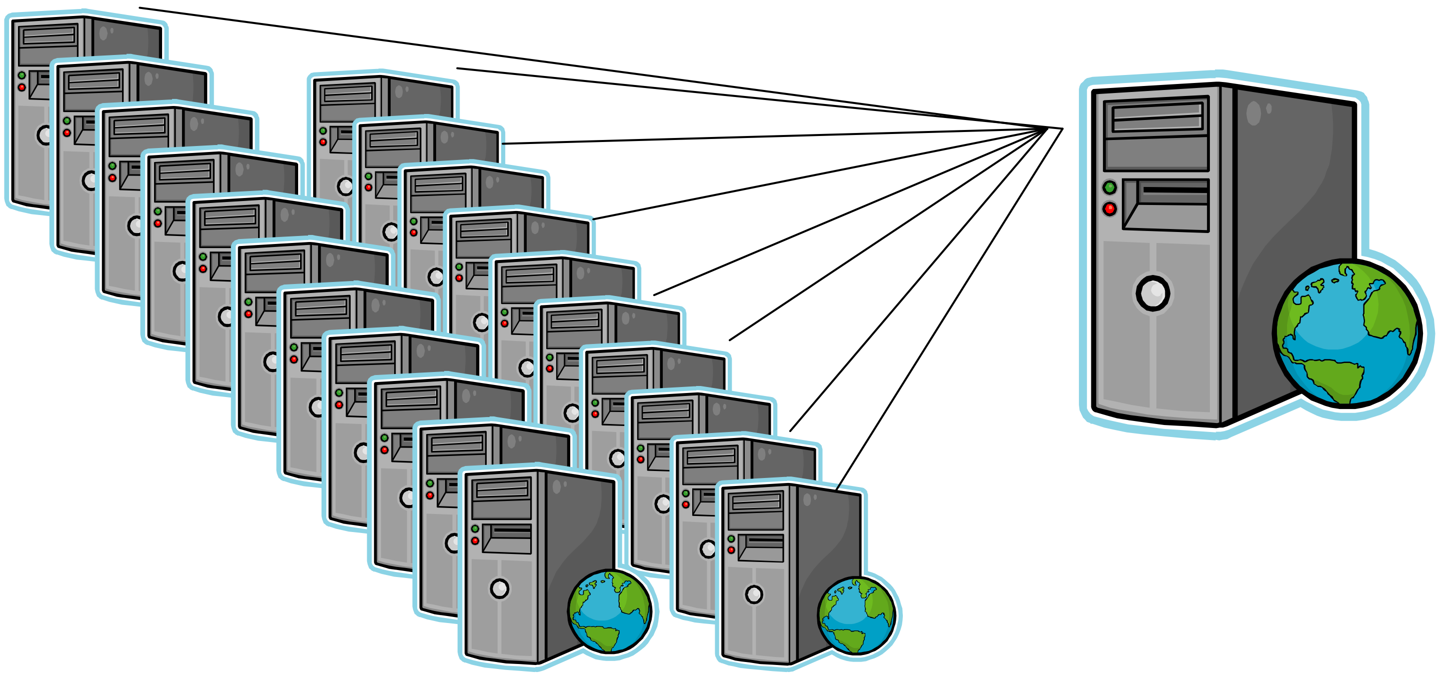
~ 10.000.000.000 comparisons ….. (10G)

→ 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

All other phonemes to 'closest sound'

Symbols for:

Nasalization
Labialization
Palatalization
Aspiration
Glottalization

Abaza (Caucasian):

Meaning

PERSON

LEAF

SKIN

HORN

NOSE

TOOTH

Abaza (Caucasian):

Meaning	IPA
PERSON	ʕʷɨʧʼʲʷʕʷɨs
LEAF	bɣʲɨ
SKIN	ʧʷazʲ
HORN	ʧʼʷɨʕʷa
NOSE	pɨnʦʼa
TOOTH	pɨʦ

Abaza (Caucasian):

Meaning	IPA		ASJPcode
PERSON	ʕ^{w}ɨʧ'^{jw}ʕ^{w}ɨs	→	Xw3Cw"yXw3s
LEAF	bɣ^{j}ɨ	→	bxy3
SKIN	ʧ^{w}az^{j}	→	Cwazy
HORN	ʧ'^{w}ɨʕ^{w}a	→	Cw"3Xwa
NOSE	pɨnʦ'a	→	p3nc"a
TOOTH	pɨʦ	→	p3c

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

C A T

D O G

x x x = 3

Comparing words

1. Value depends on length of longest word

C A T

D O G

x x x = 3

E L E P H A N T

D O G

x x x x x x x x x = 8

Comparing words

1. Value depends on length of longest word

→ Normalize: LDN = (LD / L_{max})

Comparing words

1. Value depends on length of longest word

C A T

D O G

x x x = 3/3 = 1.0

E L E P H A N T

D O G

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: LDN = (LD / L_{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**

Comparing words

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**

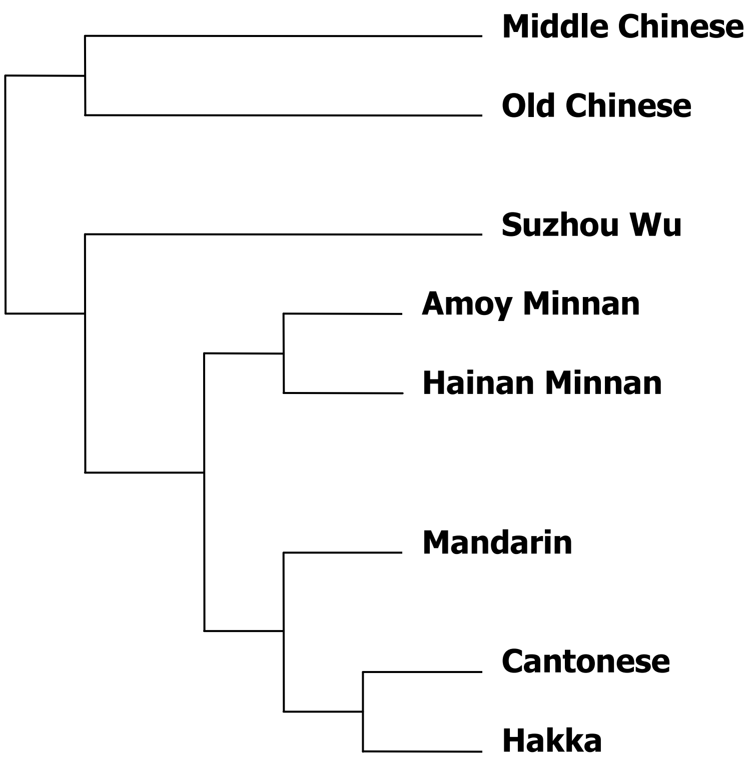
→ **Normalize: LDN = (LD / L_{max})**

2. **Differences between lgs in phonological overlap**

→ **Eliminate ‘ background noise’:**

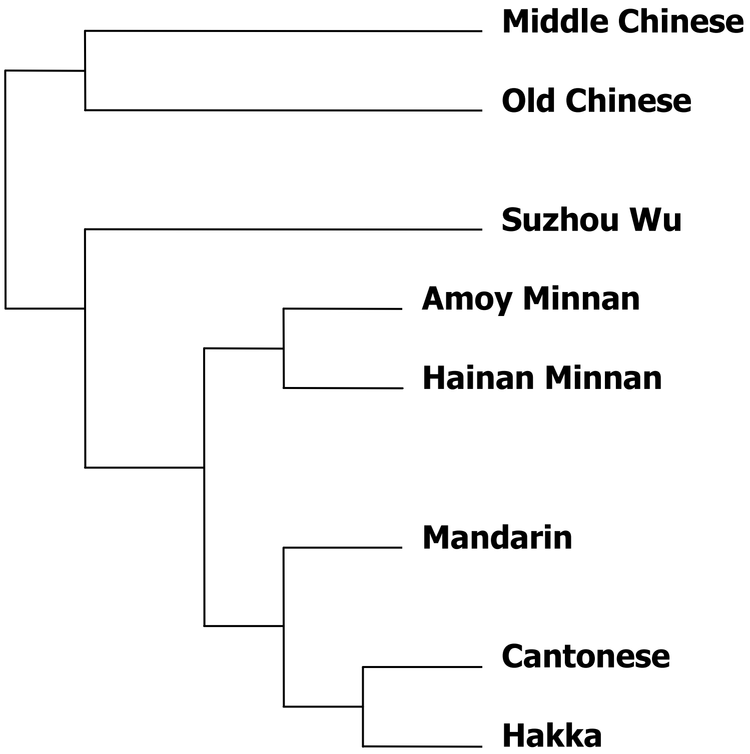
LDND = (LDN / $LDN_{different\ pairs}$)

Sino-Tibetan: Chinese



ASJP tree based on lexical relations

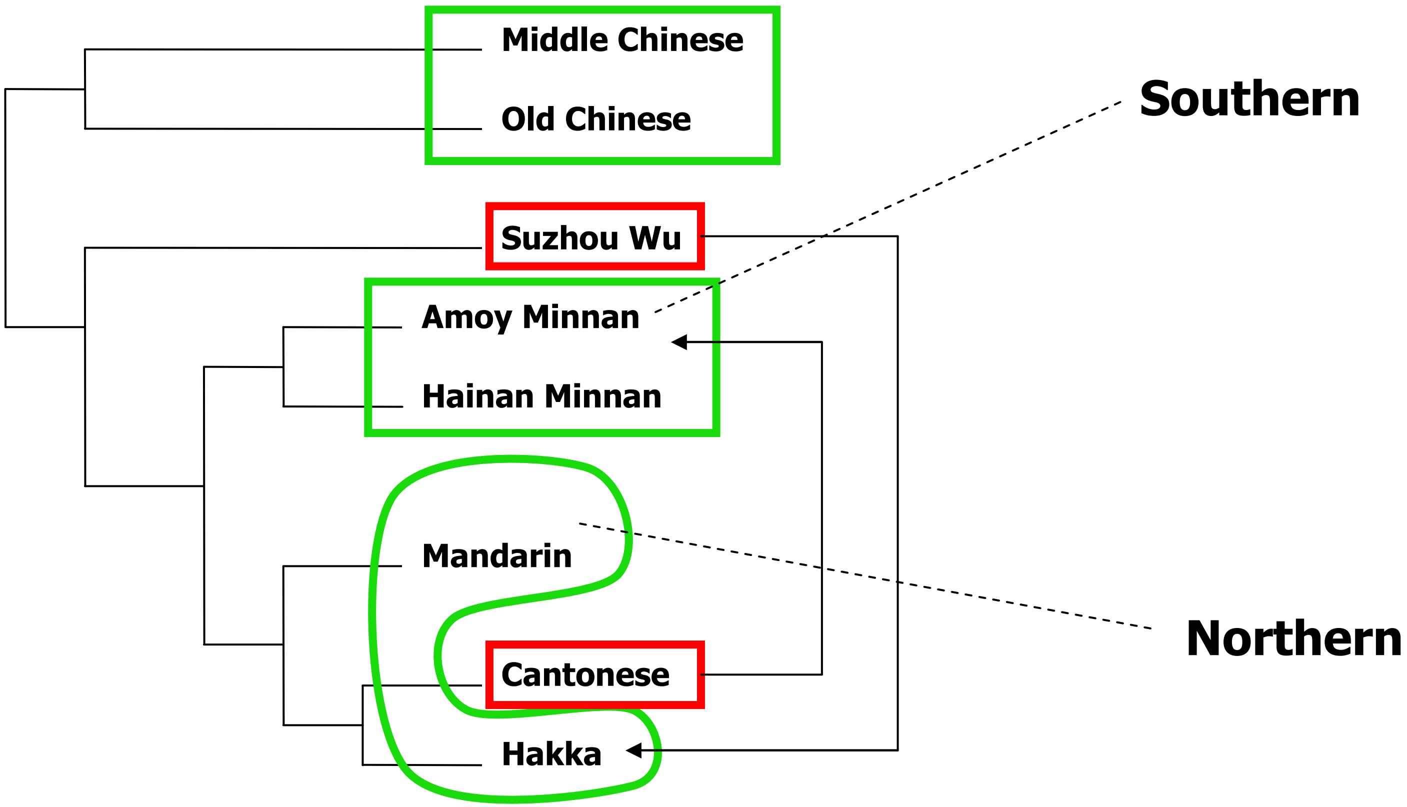
Sino-Tibetan: Chinese



ALL & ONLY

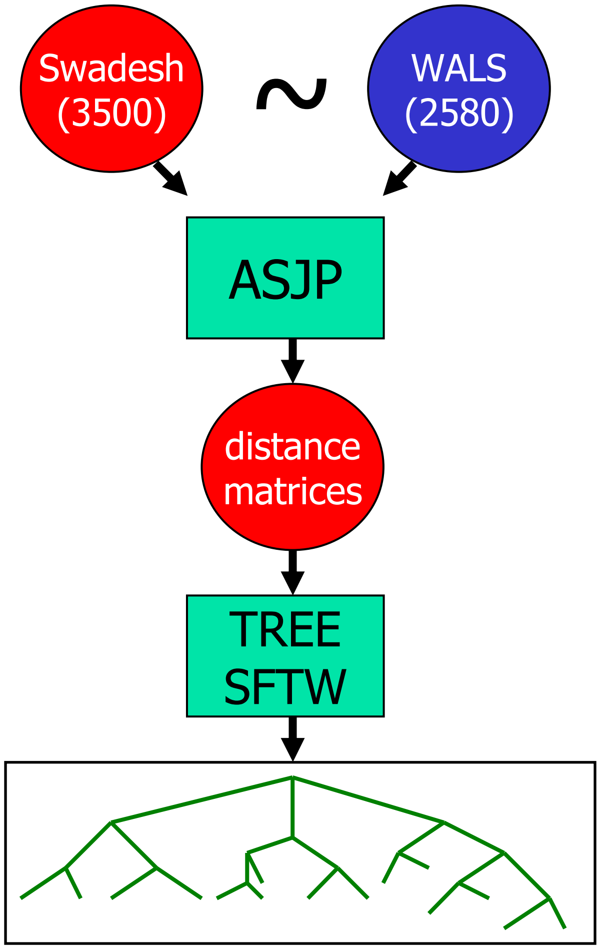
ASJP tree based on lexical relations

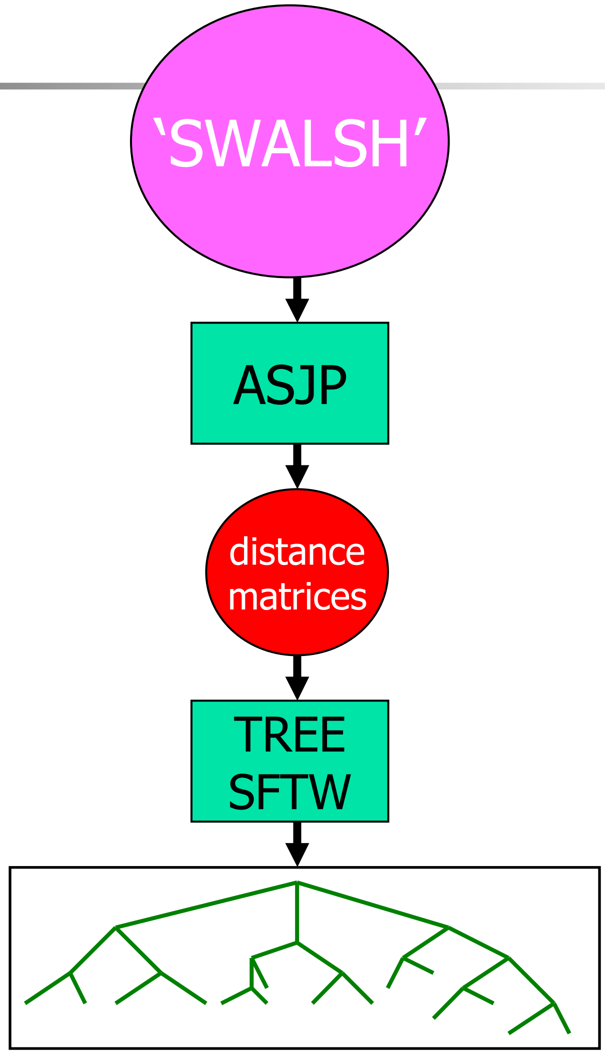
Sino-Tibetan: Chinese



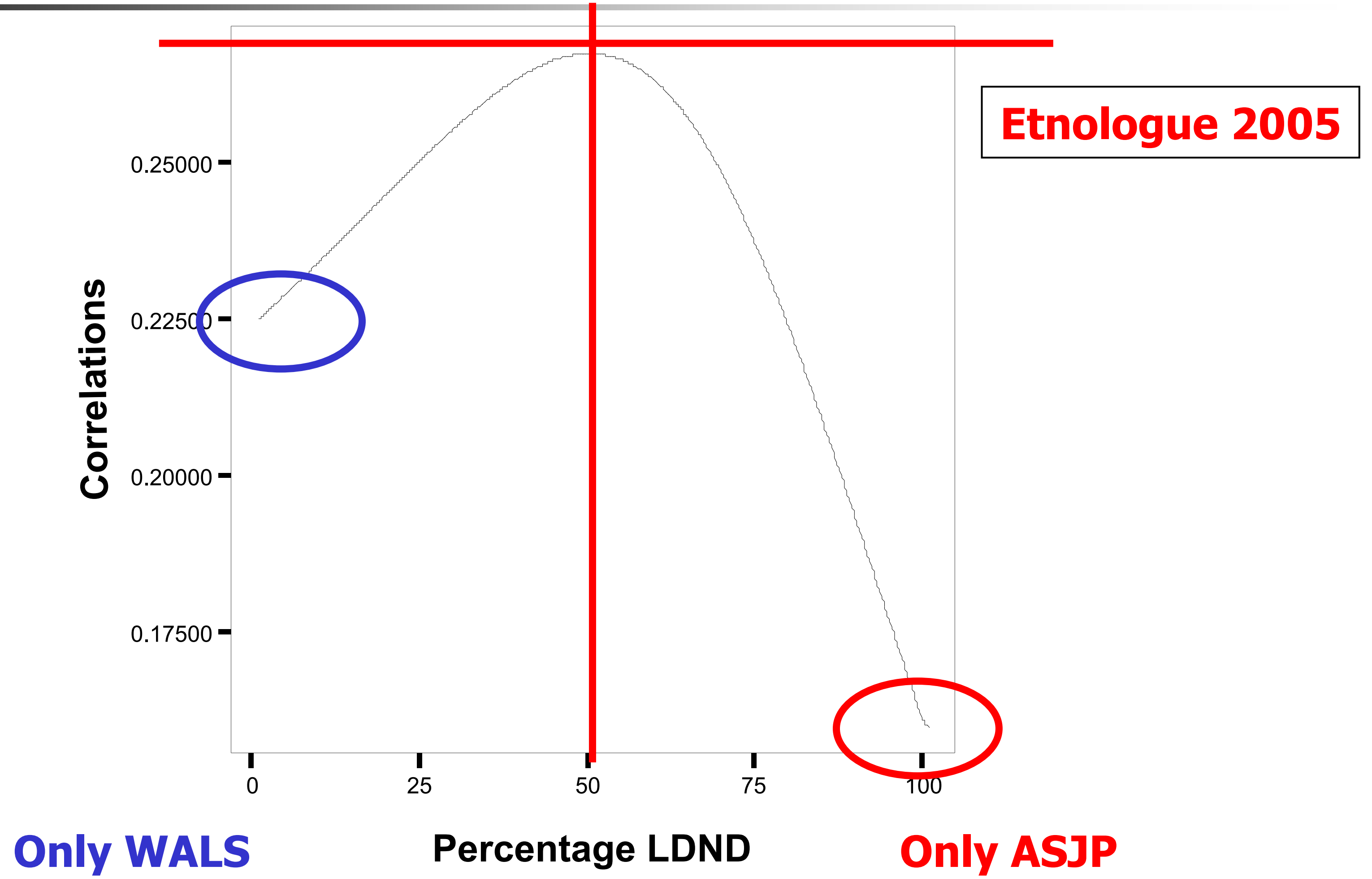
Genetic classification in Thurgood & LaPolla (eds)

Lexical plus typological data

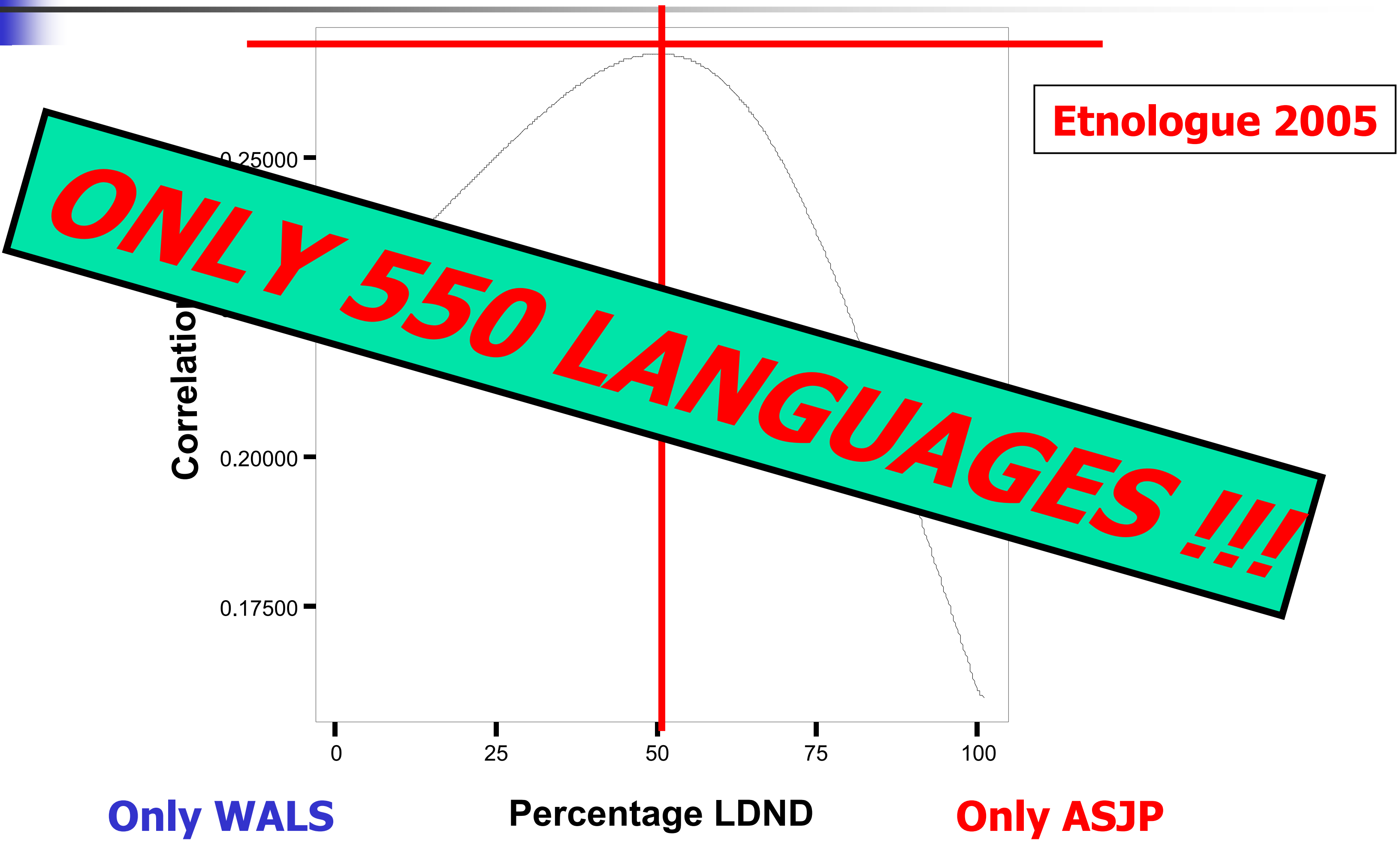



'SWALSH'
ASJP
distance matrices
TREE SFTW

Improving the fit



Improving the fit



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:

Lexical items: transcription

<u>Second phase of project (2009-10):</u>

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

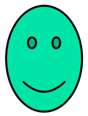
<u>Abaza (Caucasian):</u>

Meaning: PERSON

Lexical items: transcription

<u>Abaza (Caucasian):</u>

Meaning: PERSON

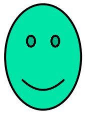


IPA: ʕʷɨʧʼʲʷʕʷɨs

Lexical items: transcription

Abaza (Caucasian):

Meaning: PERSON



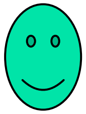
IPA: ʕʷɨʧʼʲʷʕʷɨs

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

Lexical items: transcription

Abaza (Caucasian):

Meaning: PERSON



IPA: ʕʷɨʧʼʲʷʕʷɨs

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

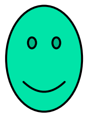
'a' <- 661, 895, 416, ...

formal grammar

Lexical items: transcription

<u>Abaza (Caucasian):</u>

Meaning: PERSON



IPA: ʕʷɨtʃ’ʲʷʕʷɨs

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

‘a’ <- 661, 895, 416, ...

formal grammar

ASJP++code

Lexical items: transcription

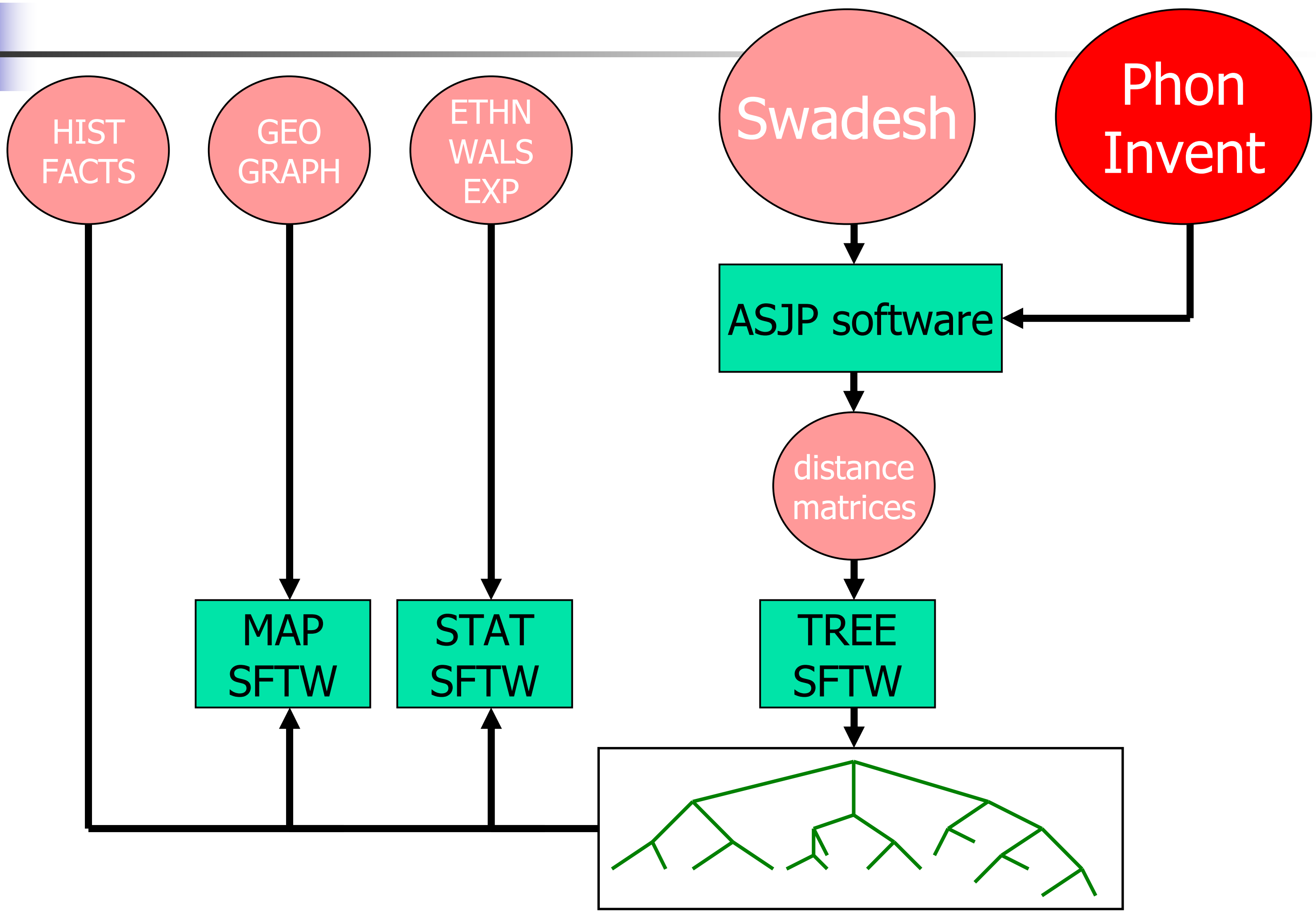
IPA: ʕʷɨʧʼʲʷʕʷɨs

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]

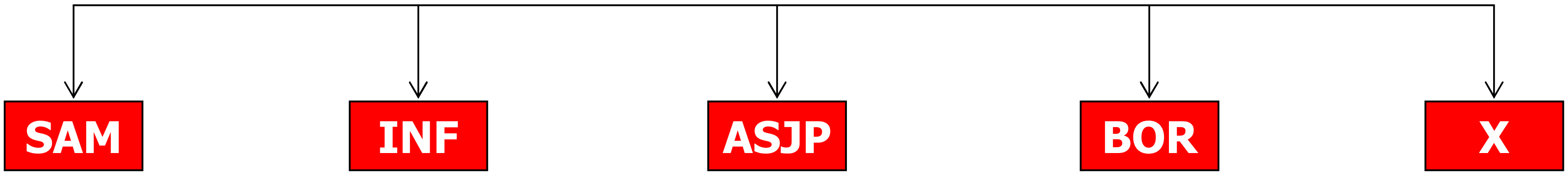
formal grammar
+
phonological features

ASJP++code: (comparison of phonological features)

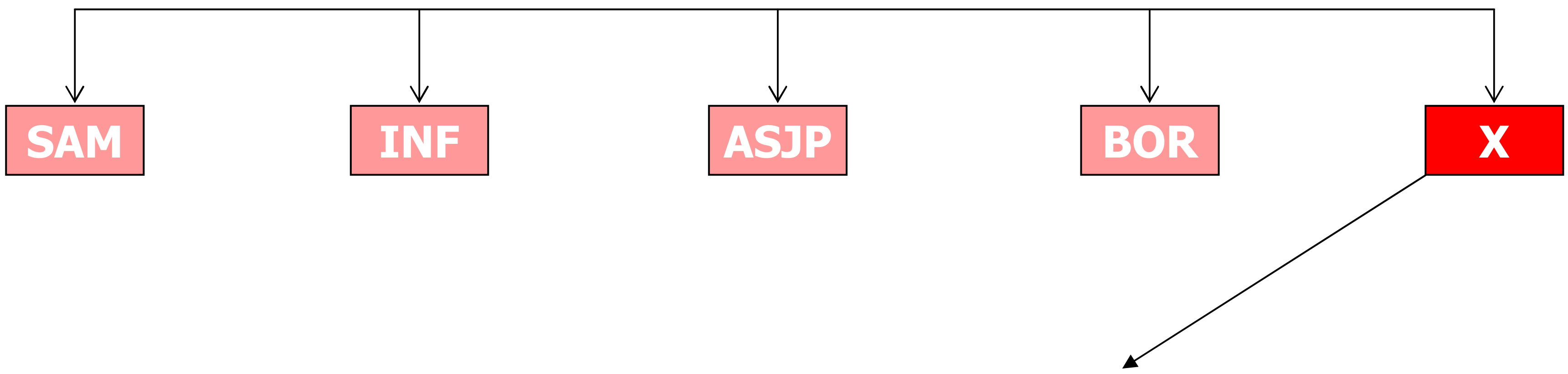

HIST
FACTS
GEO
GRAPH
ETHN
WALS
EXP
Swadesh
Phon
Invent
ASJP software
distance
matrices
MAP
SFTW
STAT
SFTW
TREE
SFTW

5. Accessibility

Accessibility

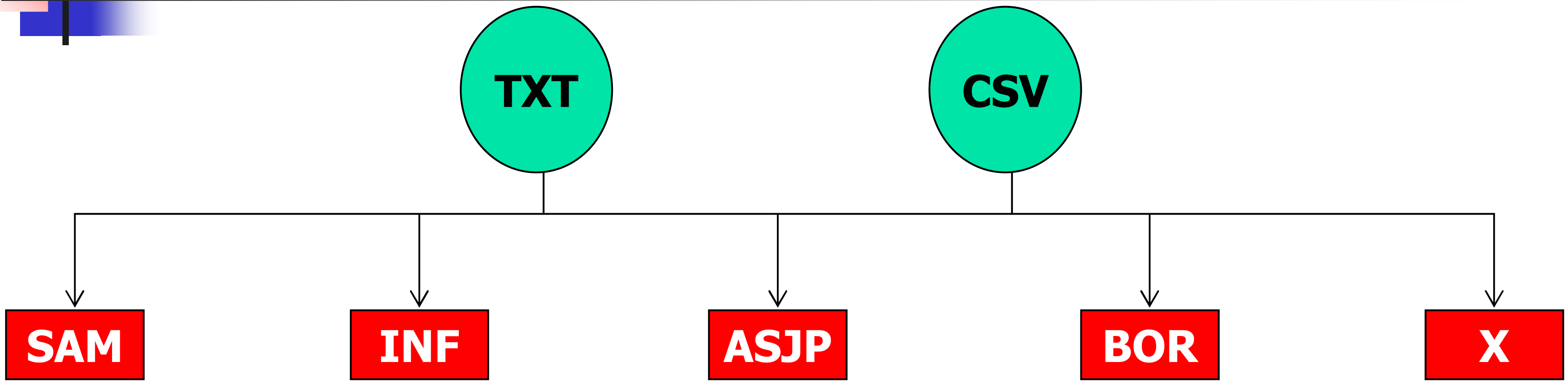


Accessibility



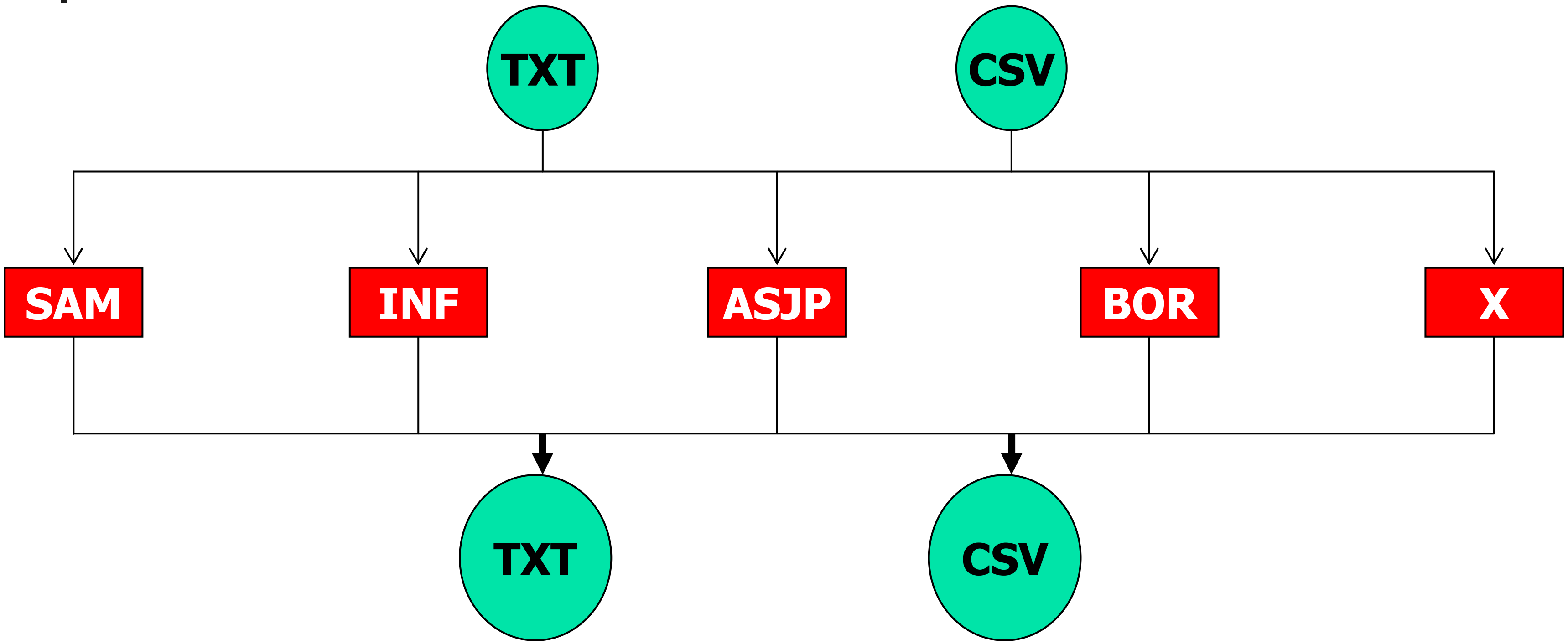
<u>Small Tools:</u>
Lookup Ethnologue code
Affiliation
Linguistic variables
...

Access: data, internal



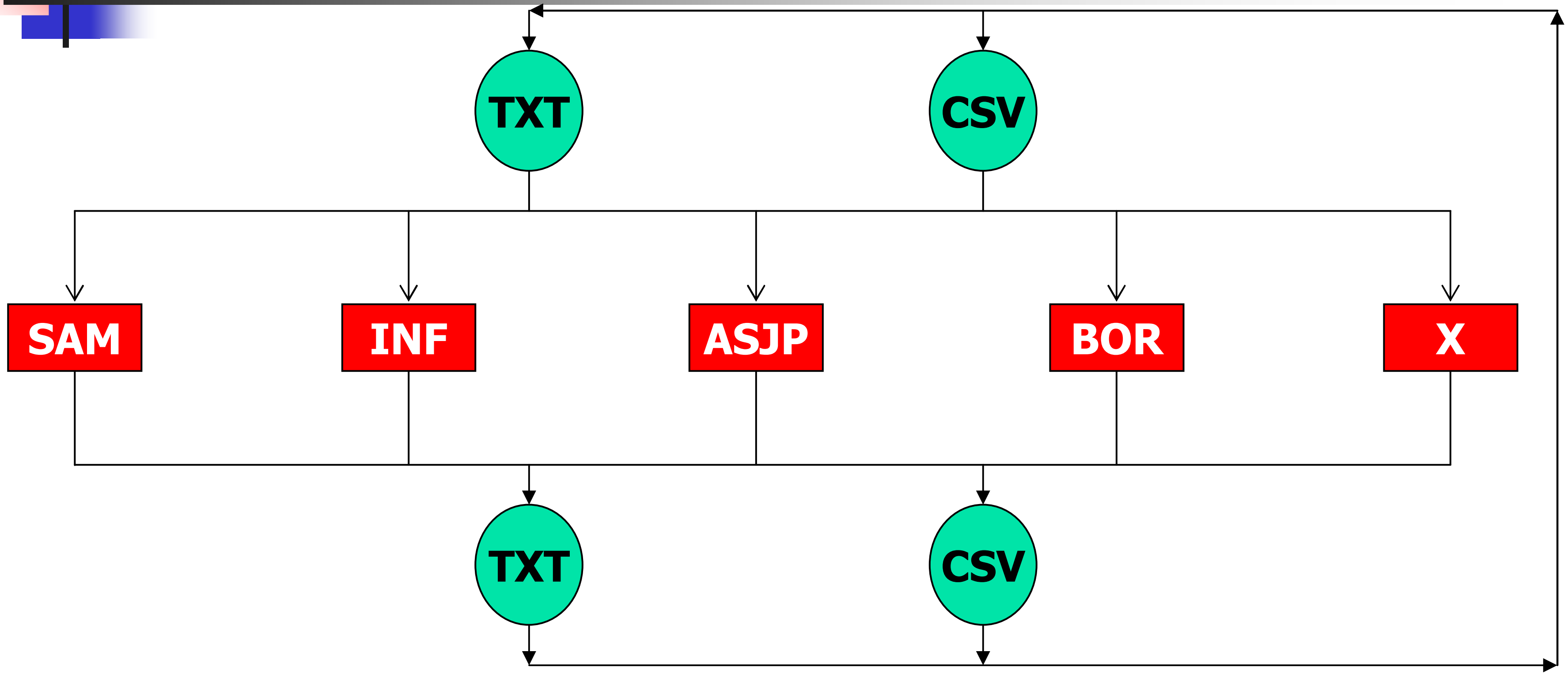
Generally accepted data structures (Unicode; UTF8)

Access: data, internal

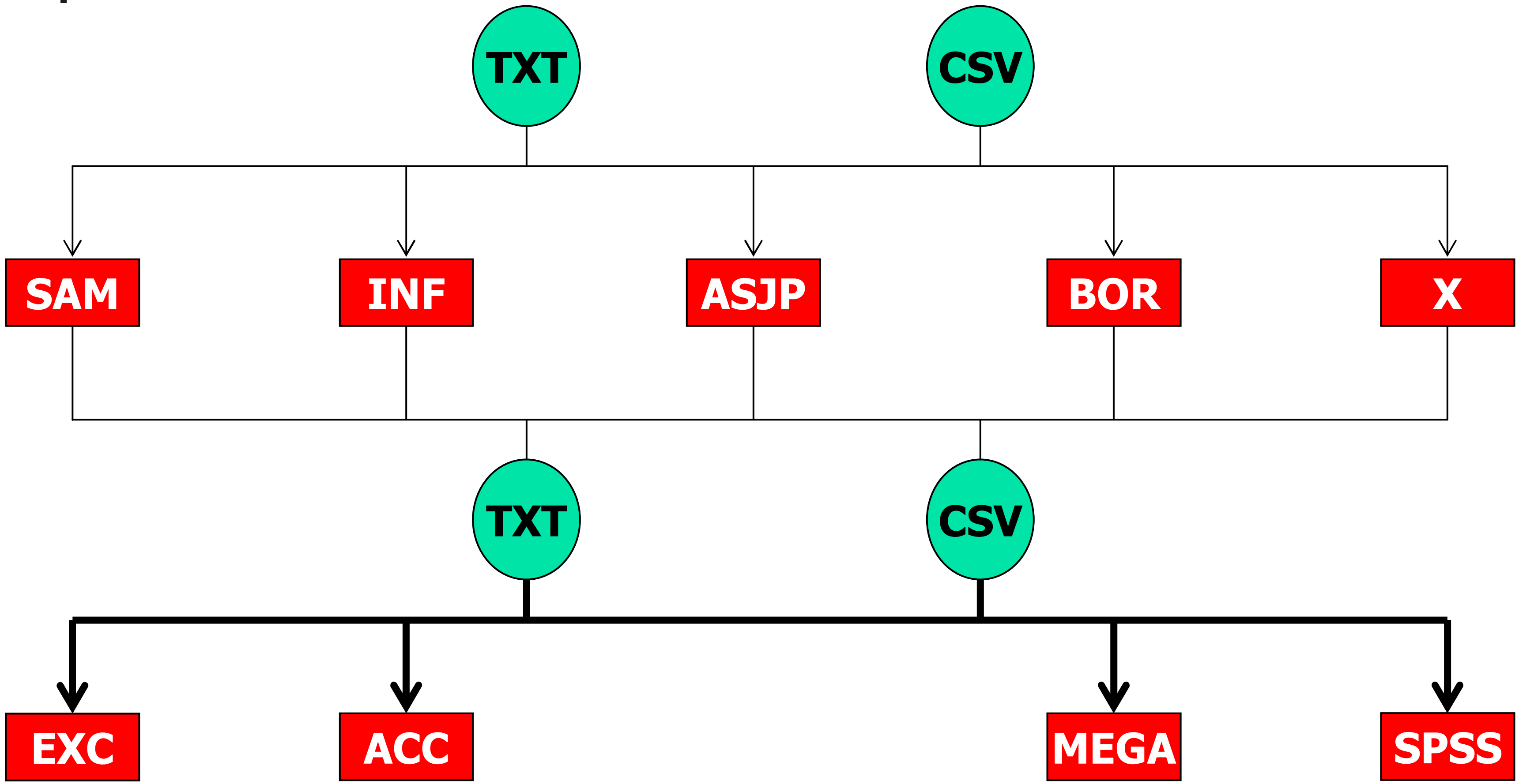


Generally accepted data structures (Unicode; UTF8)

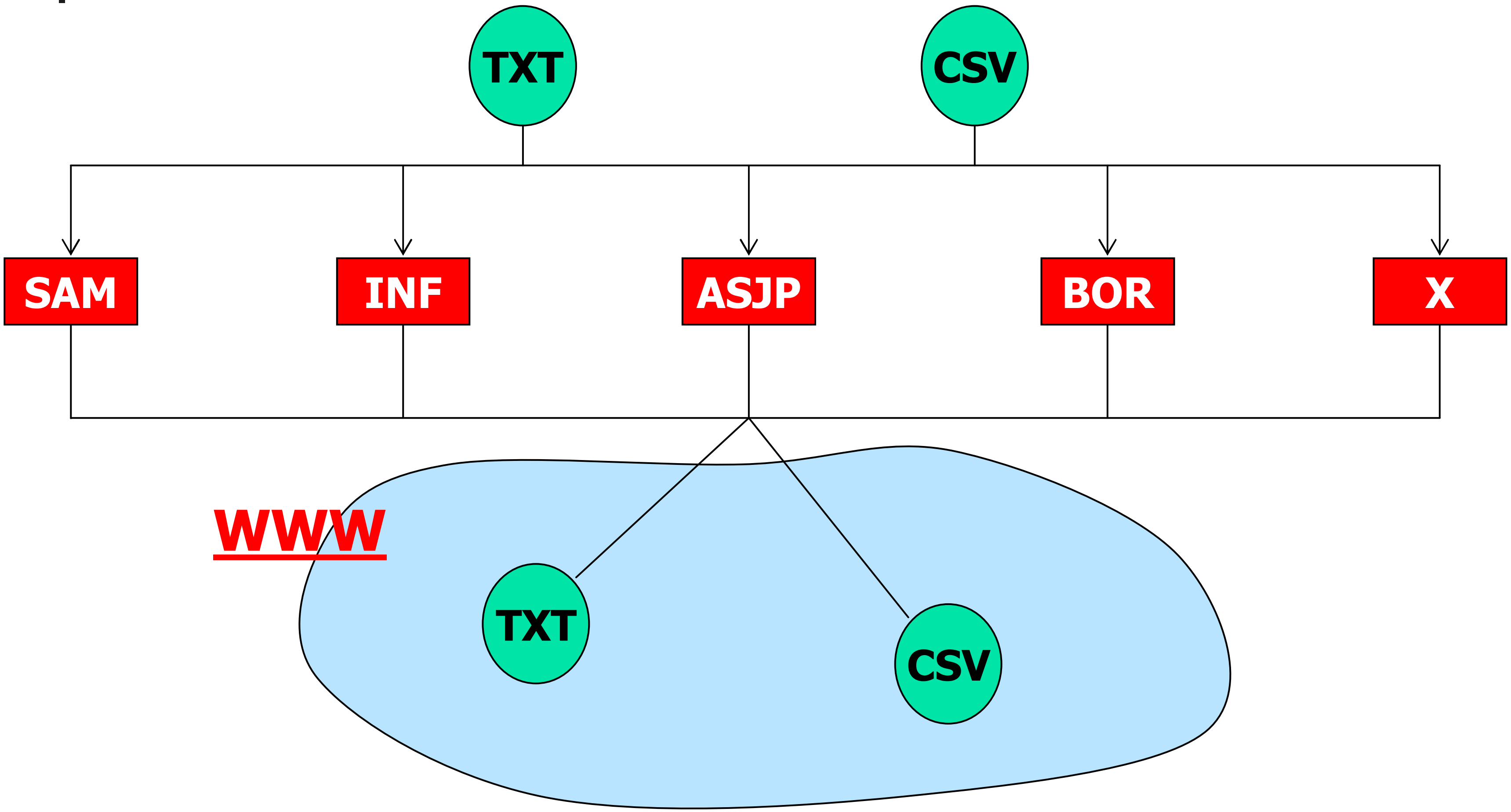
Access: data, internal



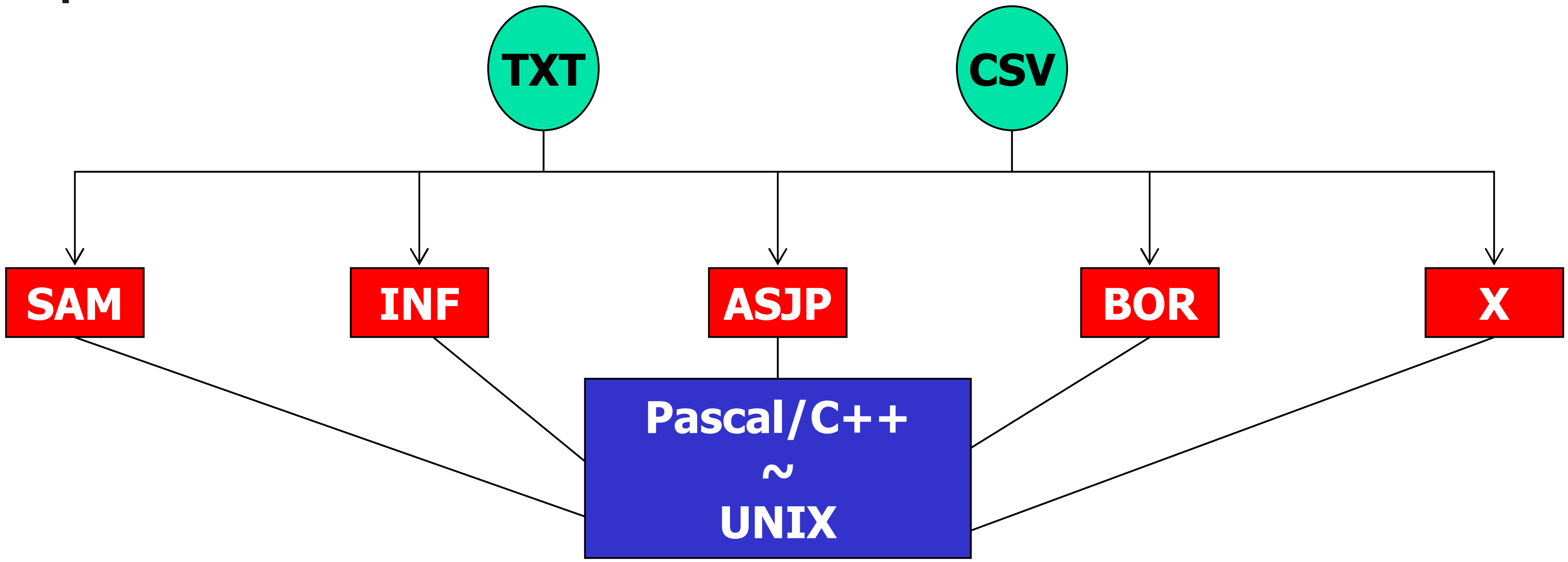
Access: data, external



Access: data, universal?

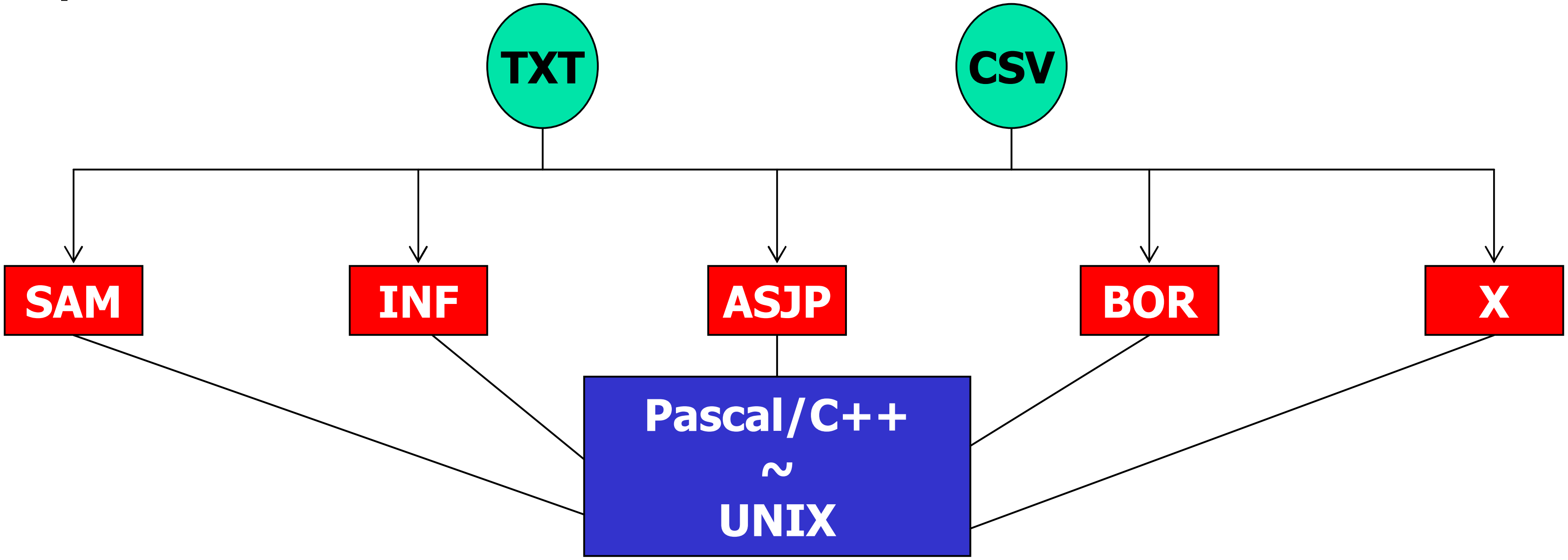


Access: software: in/external?



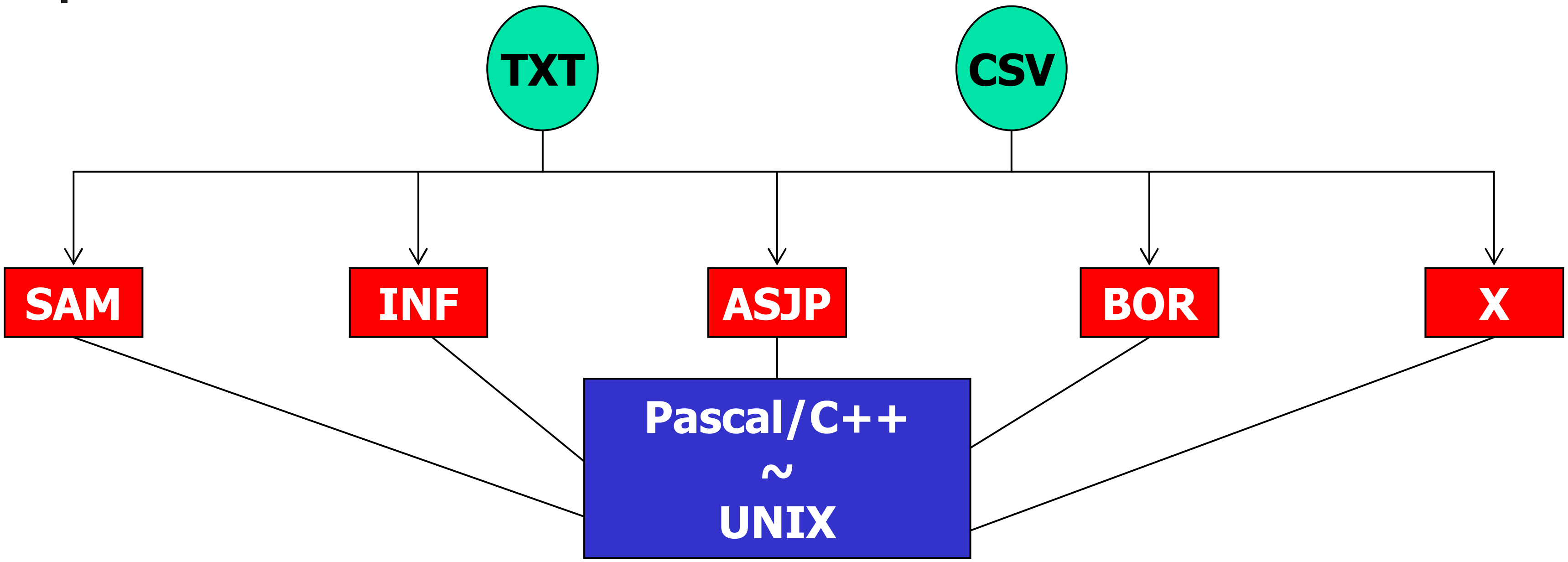
1. Too big / slow for Windows (?)

Access: software: in/external?



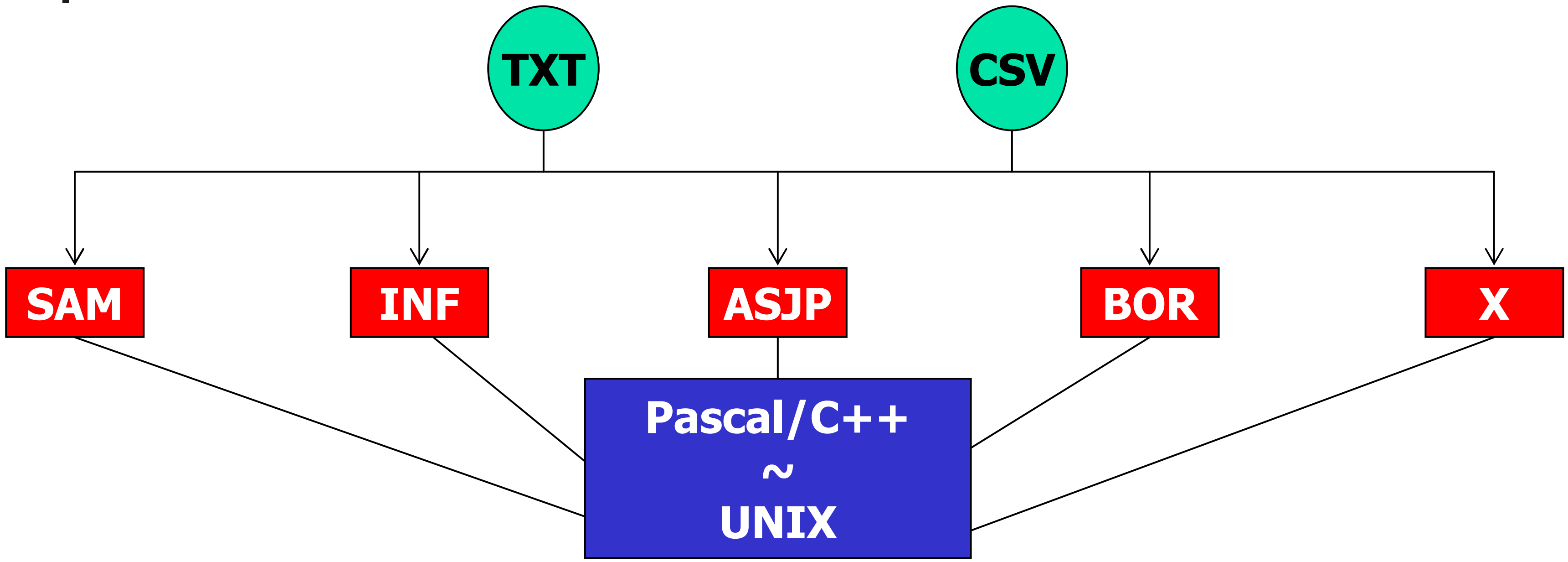
1. Too big / slow for Windows (?)
2. No user interface

Access: software: in/external?



1. Too big / slow for Windows (?)
2. No user interface

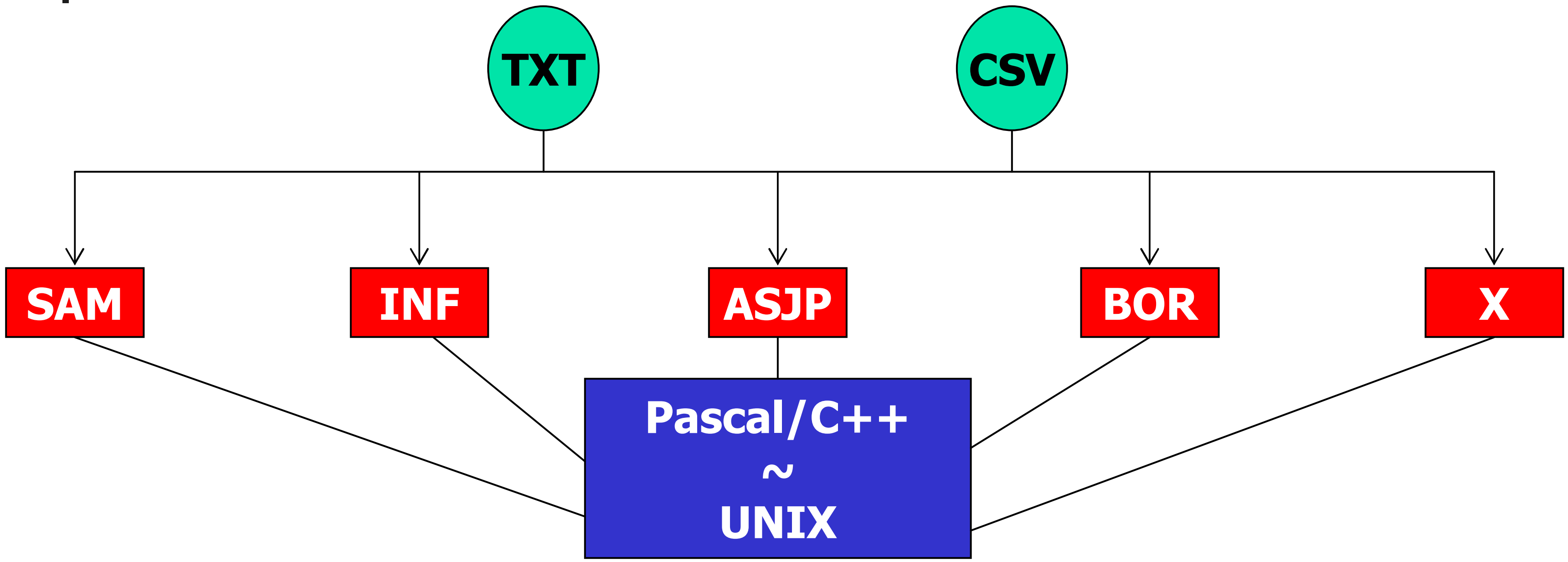
Access: software: in/external?



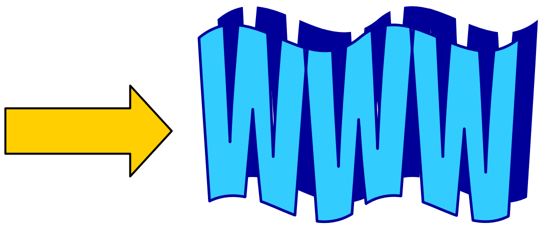
1. Too big / slow for Windows (?)
2. No user interface

→ . . . .

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

?