# Investigating the Impact of Sample Size on Cognate Detection

Johann-Mattis List

Research Unit Quantitative Language Comparison
Philipps-University Marburg

March 17, 2013

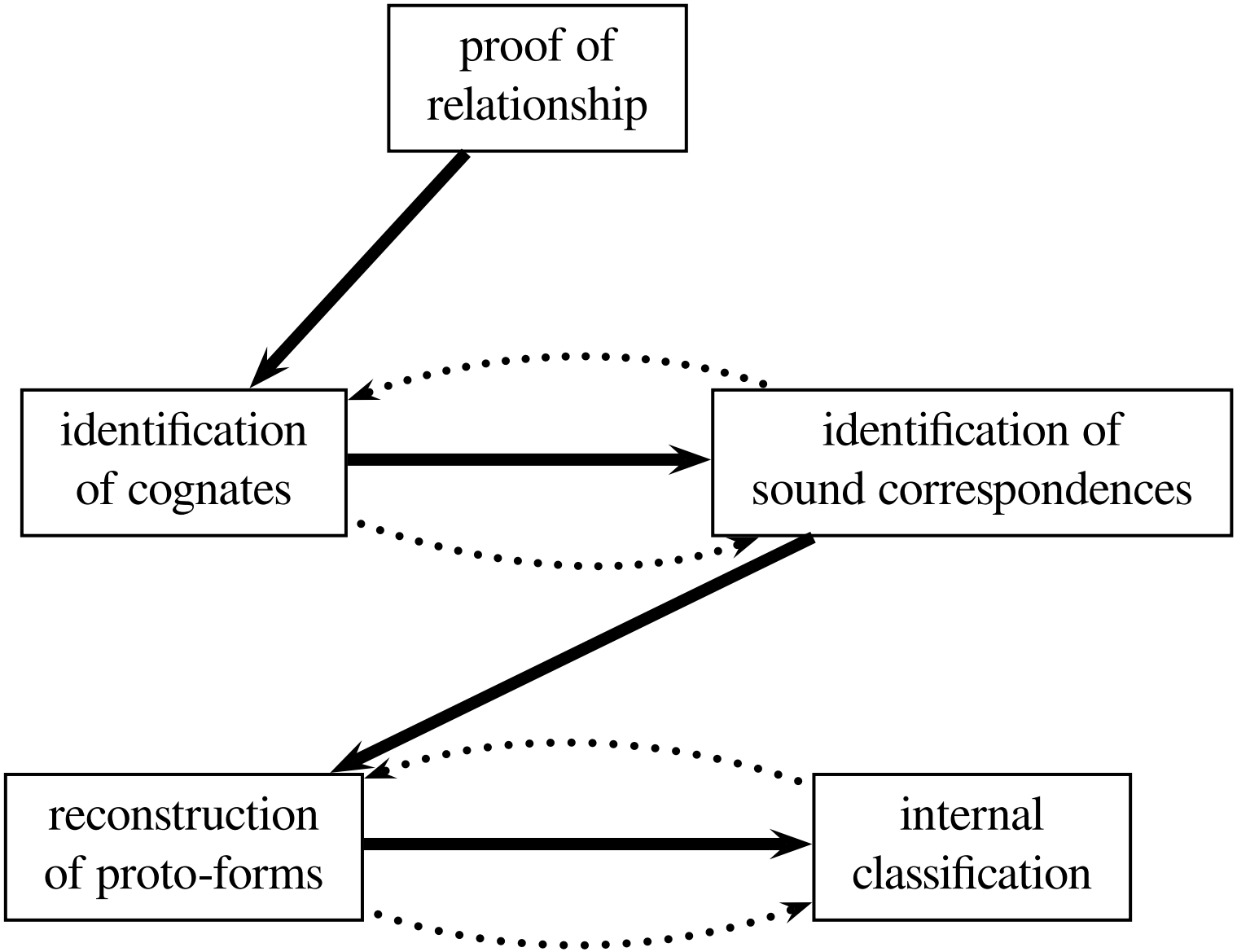
## Sanscruta and Italian

Sono scritte le loro scienze tutte in una lingua, che dimandano Sanscruta, che vuol dire bene articolata. [...] et ha la lingua d'oggi molte cose comuni con quella, nella quale sono molti de' nostri nomi, e particularmente de' numeri il 6, 7, 8 e 9, Dio, serpe, et altri assai.(Sassetti 1855: 415)

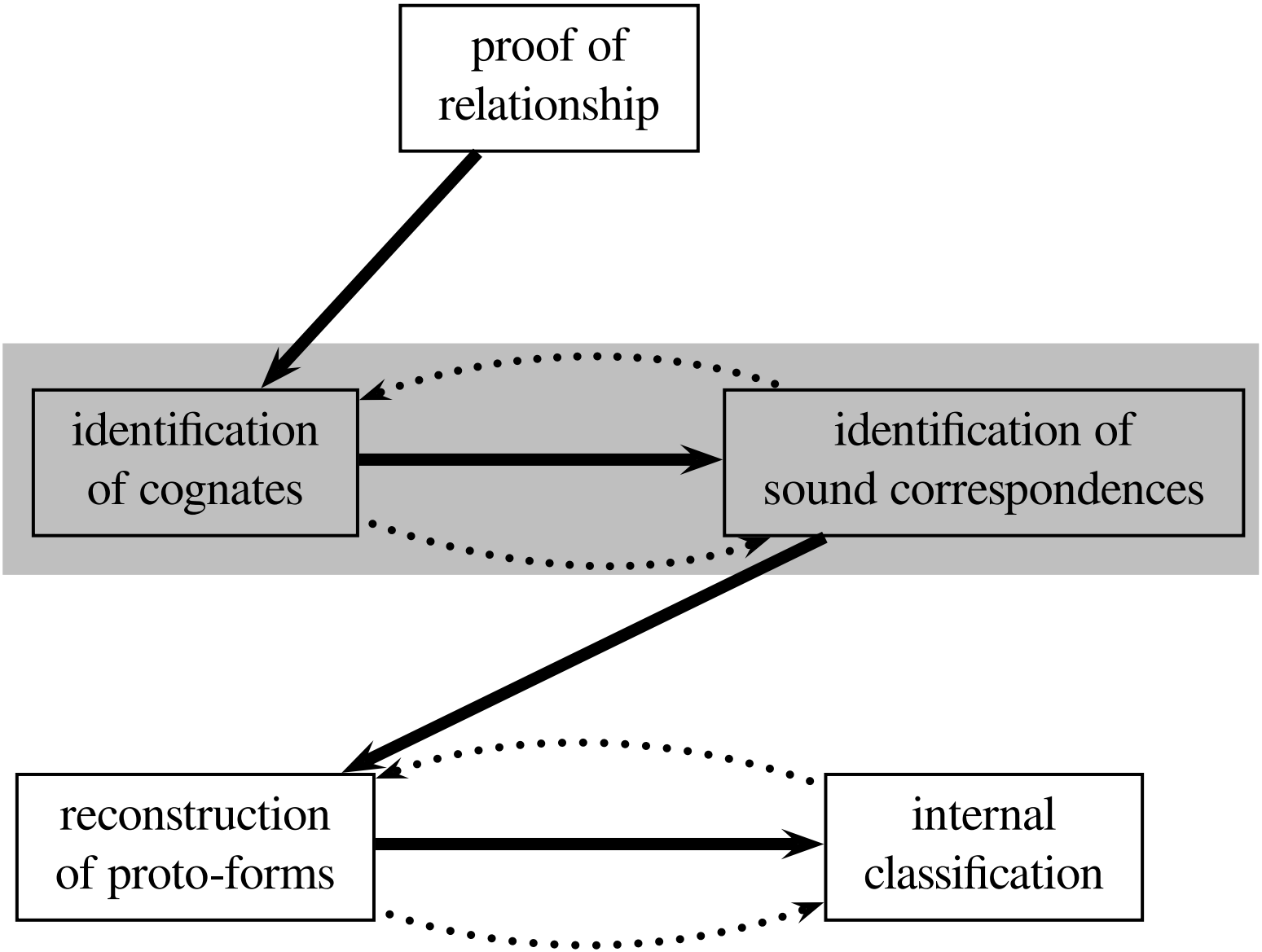
**Translation:** Everything that is related to science is written in a language which they call "Sanscruta", meaning as much as "well-articulated". Our language has much in common with it, among others many of our words, especially the numbers 6, 7 , 8, and 9, "God", "snake", and many more.

# The Comparative Method

# Working Procedure



# Working Procedure



# Cognate Detection

# Cognate Detection

| **Cognate List** | |
|---|---|
| German | *dünn* |
| English | *thin* |
| German | *Ding* |
| English | *thing* |
| German | *dumm* |
| English | *dumb* |

# Cognate Detection

| Cognate List | | Alignment | | |
|---|---|---|---|---|
| German | *dünn* | d | ʏ | n |
| English | *thin* | θ | ɪ | n |
| German | *Ding* | d | ɪ | ŋ |
| English | *thing* | θ | ɪ | ŋ |
| German | *dumm* | d | ʊ | m |
| English | *dumb* | d | ʌ | m |

# Cognate Detection

| **Cognate List** | | **Alignment** | | |
|---|---|---|---|---|
| German | *dünn* | d | ʏ | n |
| English | *thin* | θ | ɪ | n |
| German | *Ding* | d | ɪ | ŋ |
| English | *thing* | θ | ɪ | ŋ |
| German | *dumm* | d | ʊ | m |
| English | *dumb* | d | ʌ | m |

| **Correspondence List** | | |
|---|---|---|
| GER | ENG | Frequ. |
| d | θ | 2 x |
| d | d | 1 x |
| n | n | 1 x |
| m | m | 1 x |
| ŋ | ŋ | 1 x |

# Cognate Detection

| Cognate List | | Alignment | | |
|---|---|---|---|---|
| German | *dünn* | d | ʏ | n |
| English | *thin* | θ | ɪ | n |
| German | *Ding* | d | ɪ | ŋ |
| English | *thing* | θ | ɪ | ŋ |
| German | *dumm* | d | ʊ | m |
| English | *dumb* | d | ʌ | m |
| German | *Dorn* | d | ɔɐ | n |
| English | *thorn* | θ | ɔː | n |

| Correspondence List | | |
|---|---|---|
| GER | ENG | Frequ. |
| d | θ | 2 x |
| d | d | 1 x |
| n | n | 1 x |
| m | m | 1 x |
| ŋ | ŋ | 1 x |

# Cognate Detection

| Cognate List | | Alignment | | |
|---|---|---|---|---|
| German | *dünn* | d | ʏ | n |
| English | *thin* | θ | ɪ | n |
| German | *Ding* | d | ɪ | ŋ |
| English | *thing* | θ | ɪ | ŋ |
| German | *dumm* | d | ʊ | m |
| English | *dumb* | d | ʌ | m |
| German | *Dorn* | d | ɔɐ | n |
| English | *thorn* | θ | ɔː | n |

| Correspondence List | | |
|---|---|---|
| GER | ENG | Frequ. |
| d | θ | 3 x |
| d | d | 1 x ? |
| n | n | 2 x |
| m | m | 1 x |
| ŋ | ŋ | 1 x |

# Cognate Detection

| Cognate List | | Alignment | | |
|---|---|---|---|---|
| German | *dünn* | d | ʏ | n |
| English | *thin* | θ | ɪ | n |
| German | *Ding* | d | ɪ | ŋ |
| English | *thing* | θ | ɪ | ŋ |
| German | *dumm* | d | ʊ | m |
| English | *dumb* | d | ʌ | m |
| German | *Dorn* | d | ɔɐ | n |
| English | *thorn* | θ | ɔː | n |

| Correspondence List | | |
|---|---|---|
| GER | ENG | Frequ. |
| d | θ | 3 x |
| d | d | 1 x |
| n | n | 2 x |
| m | m | 1 x |
| ŋ | ŋ | 1 x |

# Cognate Detection

| Cognate List | | Alignment | | |
|---|---|---|---|---|
| German | *dünn* | d | ʏ | n |
| English | *thin* | θ | ɪ | n |
| German | *Ding* | d | ɪ | ŋ |
| English | *thing* | θ | ɪ | ŋ |
| German | *Dorn* | d | ɔɐ | n |
| English | *thorn* | θ | ɔː | n |

| Correspondence List | | |
|---|---|---|
| GER | ENG | Frequ. |
| d | θ | 3 x |
| n | n | 2 x |
| ŋ | ŋ | 1 x |

# Summary

## Important Aspects

- language-specific notion of word similarity
- regular sound correspondences
- iterative character

## Unspecified Parameters

- number of languages
- semantic similarity of the words
- size of the word lists

# Summary

## The Problem of the Sample Size

| | Albanian | English | French | German |
|---|---|---|---|---|
| Albanian | | 0.07 | 0.10 | 0.10 |
| English | 14 | | 0.23 | 0.56 |
| French | 20 | 46 | | 0.23 |
| German | 20 | 111 | 46 | |

Numbers and proportions of shared cognates in the Swadesh-200 list (Swadesh 1952), taken from Kessler (2001).

# Automatic Cognate Detection

# Two Types of Similarity

## "Phenotypic" Similarity (Lass 1997)

- based on surface resemblances of phonetic segments
- only depends on the words under comparison

## "Genotypic" Similarity (ibid.)

- based on sound-correspondences
- depends on the words and the languages under comparison

# Two Types of Similarity

German *Mund* [mʊnt]
English *mouth* [maʊθ]

# Two Types of Similarity

German *Mund* [mʊnt]
English *mouth* [maʊθ]

| German | | | English | | |
|---|---|---|---|---|---|
| *Milch* | [ mɪlç] | m | m | [ mɪlk] | *milk* |
| *rund* | [ rʊnt] | ʊ | au | [ raund] | *round* |
| *anders* | [ andərs] | n | - | [ ʌ(-)θər] | *other* |
| *südlich* | [ sytlɪç] | t | θ | [ sʌθərn] | *southern* |

# Language-Independent Approaches

## Normalized Edit Distance

- align two words and calculate their hamming distance
- normalize by dividing by the length of the longer word
- assume cognacy for distances beyond a certain threshold

## Turchin et al. (2010)

- convert two (or more) words to Dolgopolsky (1966) consonant classes
- assume cognacy if the first two classes match

# Language-Independent Approaches

German *Mund* [mʊnt]
English *mouth* [maʊθ]

# Language-Independent Approaches

German *Mund* [mʊnt]
English *mouth* [maʊθ]

| Turchin | | | | | NED | | | | |
|---|---|---|---|---|---|---|---|---|---|
| mʊnt | → | M | N | T | m | ʊ | n | t | |
| maʊθ | → | M | T | | m | aʊ | - | θ | |
| Matches: | | x | | | 0 | 1 | 1 | 1 | |
| 1 match => not cognate | | | | | $3/4 = 0.75$ => not cognate | | | | |

# Language-Specific Approaches

## LexStat (List 2012a)

- represent words as tuples of sound classes and prosodic strings
- use the SCA approach (List 2012b) to guess initial correspondences
- use a Monte-Carlo permutation test to derive language-specific similarity scores
- use the language-specific scores to calculate distance between words
- cluster words into cognate sets using a flat cluster algorithm

# LexStat

# LexStat

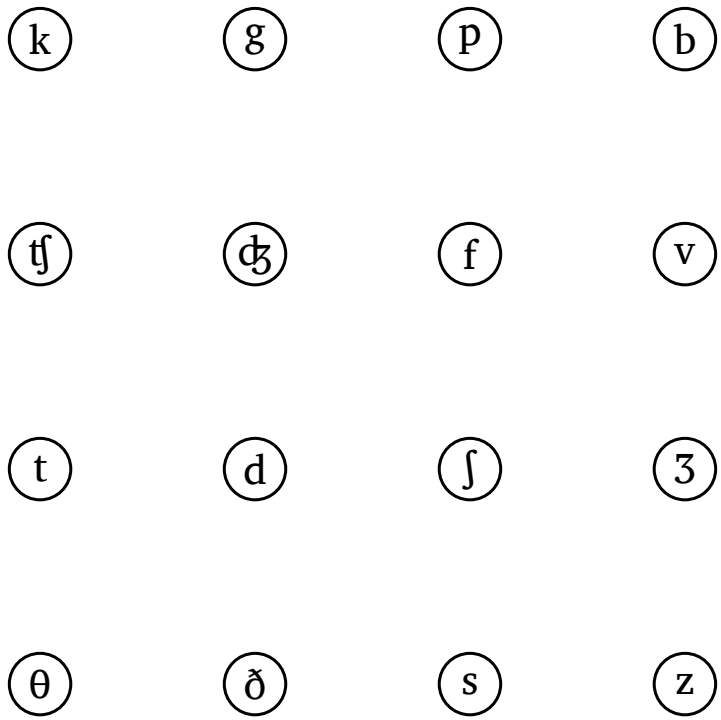
## Sound Classes

Sounds which frequently occur in correspondence relations in genetically related languages can be divided in classes (types). It is thereby assumed that “phonetic correspondences inside a ‘type’ are more regular than those between different ‘types’” (Dolgoposky 1986[1966]: 35).

# LexStat

## Sound Classes

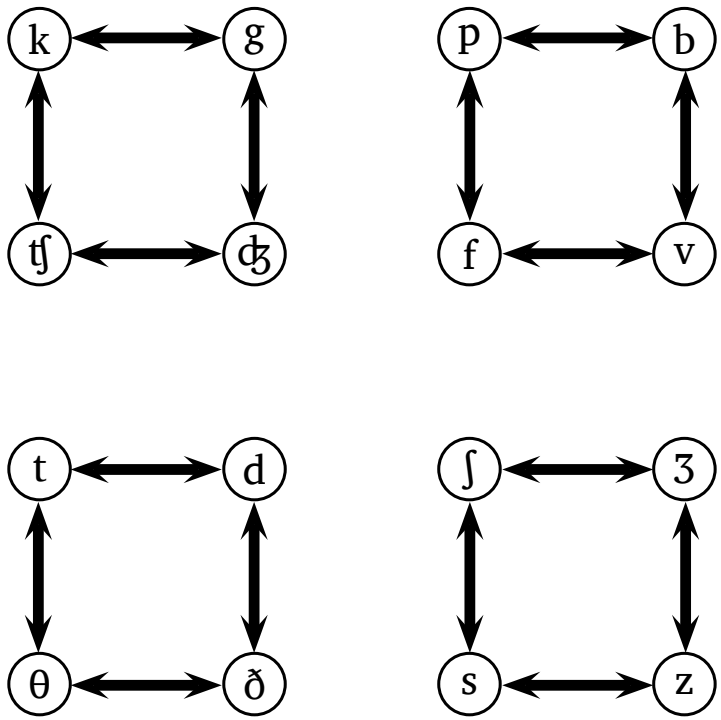
Sounds which frequently occur in correspondence relations in genetically related languages can be divided in classes (types). It is thereby assumed that “phonetic correspondences inside a ‘type’ are more regular than those between different ‘types’” (Dolgoposky 1986[1966]: 35).



# LexStat

## Sound Classes

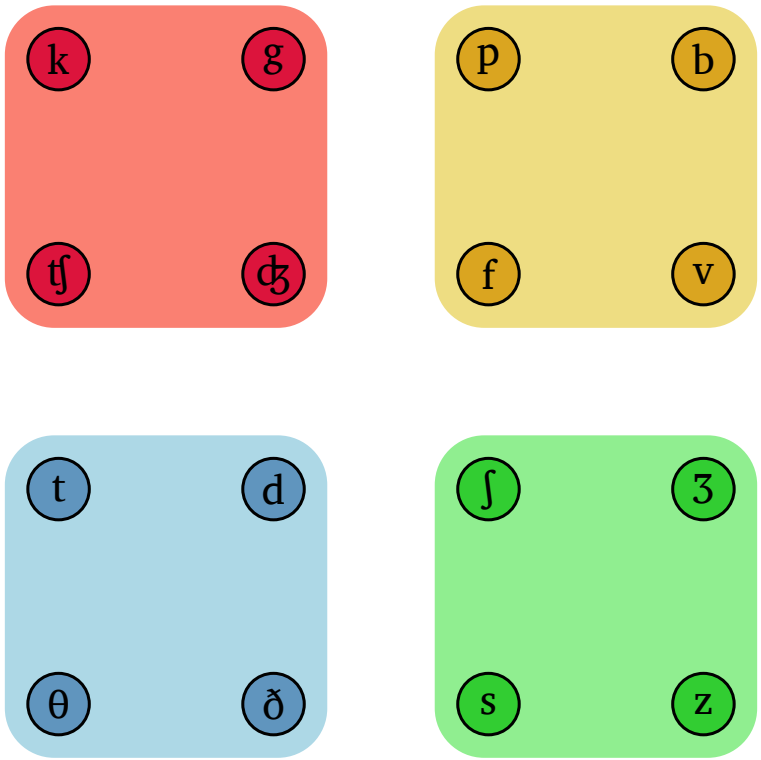
Sounds which frequently occur in correspondence relations in genetically related languages can be divided in classes (types). It is thereby assumed that "phonetic correspondences inside a 'type' are more regular than those between different 'types'" (Dolgoposky 1986[1966]: 35).



# LexStat

## Sound Classes

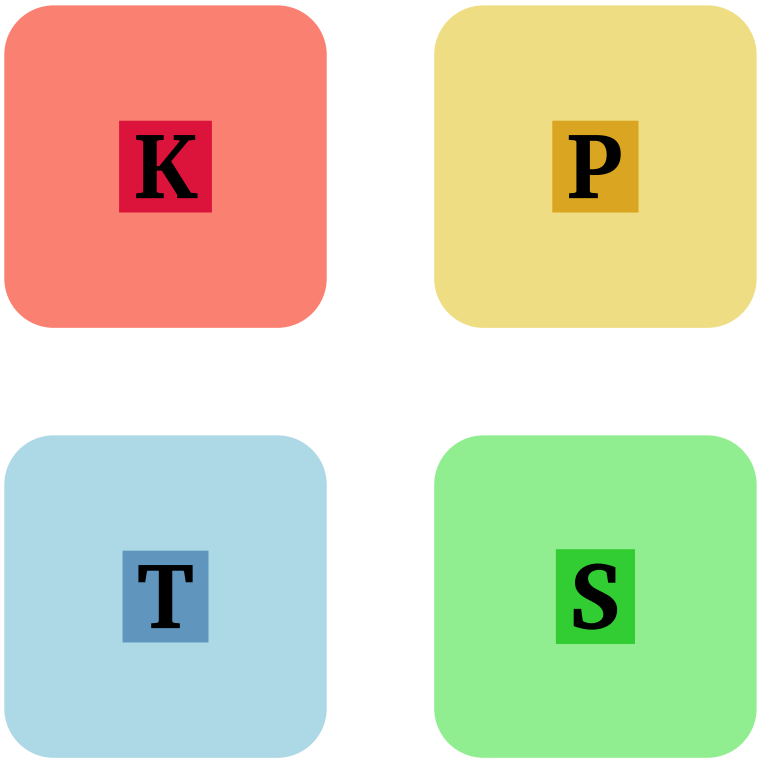
Sounds which frequently occur in correspondence relations in genetically related languages can be divided in classes (types). It is thereby assumed that “phonetic correspondences inside a ‘type’ are more regular than those between different ‘types’” (Dolgoposky 1986[1966]: 35).



# LexStat

## Sound Classes

Sounds which frequently occur in correspondence relations in genetically related languages can be divided in classes (types). It is thereby assumed that “phonetic correspondences inside a ‘type’ are more regular than those between different ‘types’” (Dolgoposky 1986[1966]: 35).



# LexStat

# LexStat

## Prosodic Strings

Sound change occurs more frequently in weak positions of sound sequences (Geisler 1992). Based on a sonority profile of sound sequences, one can distinguish sound positions according to their prosodic contexts. Prosodic context can be modeled as prosodic string in which different contexts are coded by different symbols.

# LexStat

## Prosodic Strings

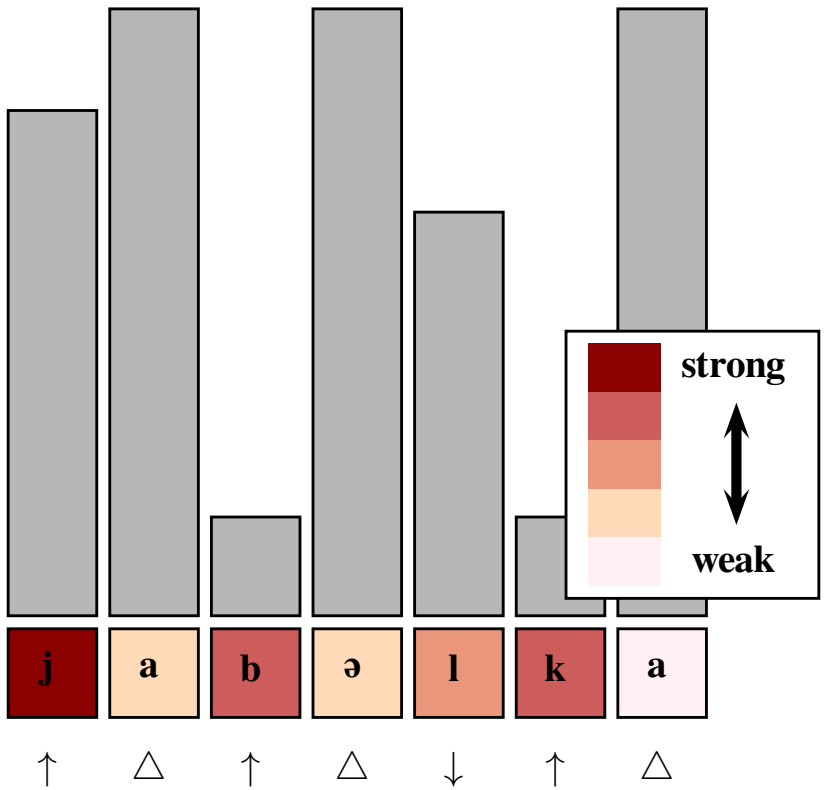
Sound change occurs more frequently in weak positions of sound sequences (Geisler 1992). Based on a sonority profile of sound sequences, one can distinguish sound positions according to their prosodic contexts. Prosodic context can be modeled as prosodic string in which different contexts are coded by different symbols.

j a b ə l k a

# LexStat

## Prosodic Strings

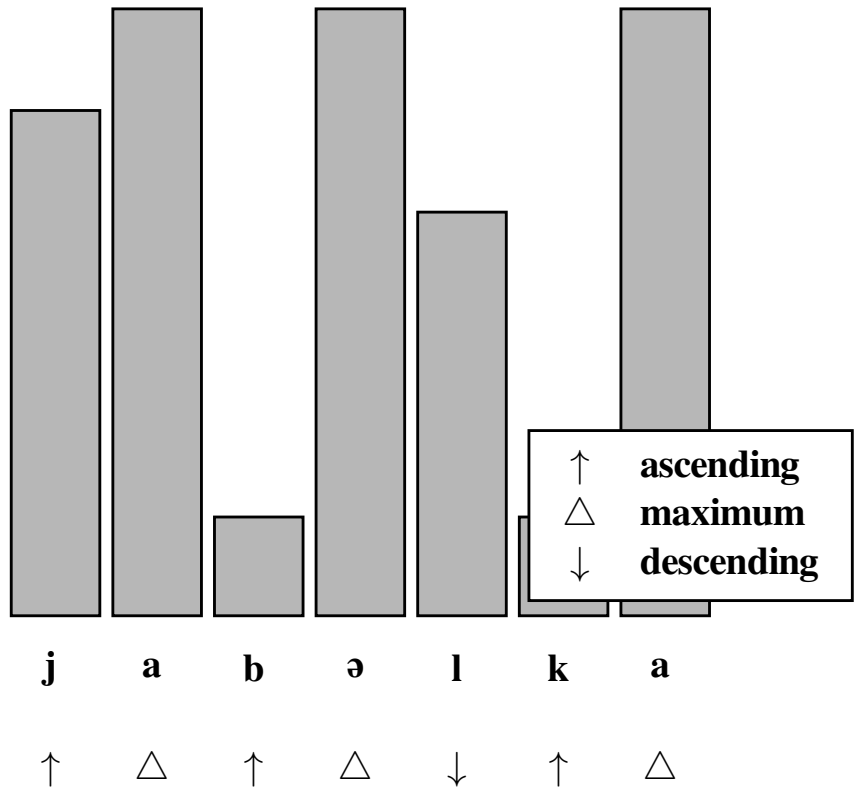
Sound change occurs more frequently in weak positions of sound sequences (Geisler 1992). Based on a sonority profile of sound sequences, one can distinguish sound positions according to their prosodic contexts. Prosodic context can be modeled as prosodic string in which different contexts are coded by different symbols.



# LexStat

## Prosodic Strings

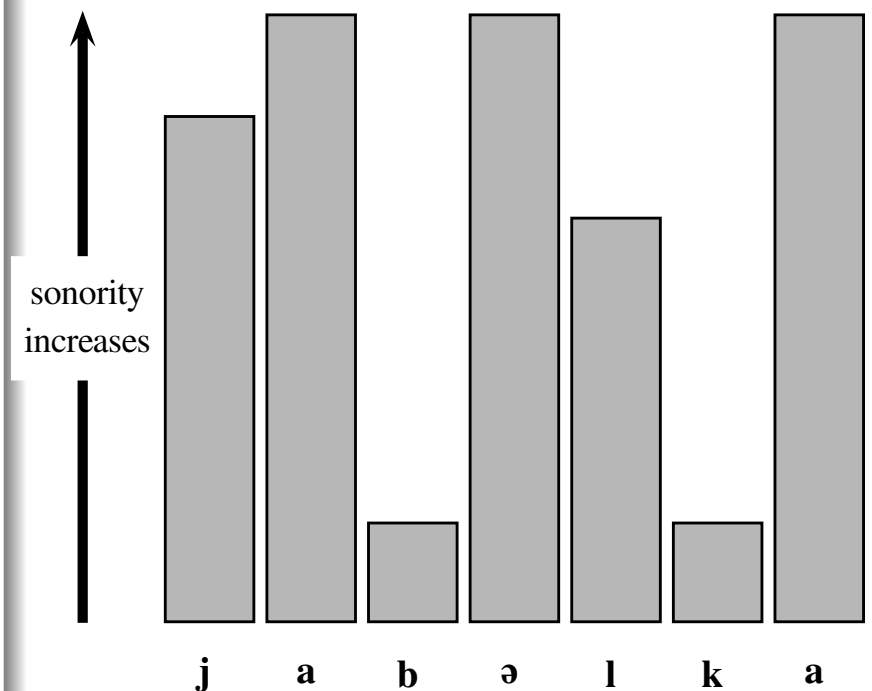
Sound change occurs more frequently in weak positions of sound sequences (Geisler 1992). Based on a sonority profile of sound sequences, one can distinguish sound positions according to their prosodic contexts. Prosodic context can be modeled as prosodic string in which different contexts are coded by different symbols.



# LexStat

## Prosodic Strings

Sound change occurs more frequently in weak positions of sound sequences (Geisler 1992). Based on a sonority profile of sound sequences, one can distinguish sound positions according to their prosodic contexts. Prosodic context can be modeled as prosodic string in which different contexts are coded by different symbols.



# LexStat

## Prosodic Strings

Sound change occurs more frequently in weak positions of sound sequences (Geisler 1992). Based on a sonority profile of sound sequences, one can distinguish sound positions according to their prosodic contexts. Prosodic context can be modeled as prosodic string in which different contexts are coded by different symbols.

| | | | | | | |
|---|---|---|---|---|---|---|
| **j** | **a** | **b** | **ə** | **l** | **k** | **a** |
| # | V | C | V | C | C | > |

# LexStat

# LexStat

| External Representation | | | | | | | |
|---|---|---|---|---|---|---|---|
| **IPA** | j | a | b | ə | l | k | a |

| Internal Representation | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Sound-Class String** | J | A | P | E | L | K | A |
| **Prosodic String** | # | V | C | V | c | C | > |

# LexStat

# LexStat

| Cognate List | | Alignment | | | |
|---|---|---|---|---|---|
| German | *Zunge* | ts | ʊ | ŋ | ə |
| English | *tongue* | t | ʌ | ŋ | - |
| German | *Zahn* | ts | aː | n | - |
| English | *tooth* | t | ʊː | - | θ |
| German | *heiß* | h | ai | s | |
| English | *hot* | h | ɔ | t | |
| German | *Fuß* | f | uː | s | |
| English | *foot* | f | ʊ | t | |

| Correspondence List | | |
|---|---|---|
| GER | ENG | Frequ. |
| ts | t | 2 x |
| s | t | 2 x |
| h | h | 1 x |
| f | f | 1 x |
| n | - | 1 x |
| … | … | … |

# LexStat

| Cognate List | | Alignment | | | | Correspondence List | | |
|---|---|---|---|---|---|---|---|---|
| German | *Zunge* | ts | ʊ | ŋ | ə | GER | ENG | Frequ. |
| English | *tongue* | t | ʌ | ŋ | - | ts | t | 2 x |
| German | *Zahn* | ts | aː | n | - | s | t | 2 x |
| English | *tooth* | t | ʊː | - | θ | h | h | 1 x |
| German | *heiß* | h | ai | s | | f | f | 1 x |
| English | *hot* | h | ɔ | t | | n | - | 1 x |
| German | *Fuß* | f | u ː | s | | … | … | … |
| English | *foot* | f | ʊ | t | | | | |

# LexStat

| Cognate List | | Alignment | | | |
|---|---|---|---|---|---|
| German | *Zunge* | C | U | N | E |
| English | *tongue* | T | A | N | – |
| German | *Zahn* | C | A | N | – |
| English | *tooth* | T | U | – | T |
| German | *heiß* | H | A | S | |
| English | *hot* | H | O | T | |
| German | *Fuß* | B | U | S | |
| English | *foot* | B | U | T | |

| Correspondence List | | |
|---|---|---|
| GER | ENG | Frequ. |
| C/# | T/# | 2 x |
| S/$ | T/$ | 2 x |
| H/$ | H/# | 1 x |
| B/$ | B/# | 1 x |
| N/c | – | 1 x |
| ... | ... | ... |

# LexStat

| “to dig” (30) | | | Turchin | NED | LexStat |
|---|---|---|---|---|---|
| Albanisch | *gërmon* | gərmo | 1 | 1 | 1 |
| Englisch | *digs* | dɪg | 2 | 2 | 2 |
| Französisch | *creuse* | krøze | 1 | 3 | 3 |
| Deutsch | *gräbt* | graːb | 1 | 1 | 4 |
| Hawaii | *ʻeli* | ʔeli | 5 | 5 | 5 |
| Navajo | *hahashgéédd* | hahageːd | 6 | 6 | 6 |
| Türkisch | *kazıyor* | kaz | 7 | 3 | 7 |

Dataset of Kessler (2001)

# LexStat

| “mouth” (104) | | | Turchin | NED | LexStat |
|---|---|---|---|---|---|
| Albanisch | *gojë* | goj | 1 | 1 | 1 |
| Englisch | *mouth* | mauθ | 2 | 2 | 2 |
| Französisch | *bouche* | buʃ | 3 | 3 | 3 |
| Deutsch | *Mund* | mund | 4 | 4 | 2 |
| Hawaii | *waha* | waha | 5 | 5 | 5 |
| Navajo | *ʾazéé́ʾ* | zeːʔ | 6 | 6 | 6 |
| Türkisch | *ağız* | aɣz | 7 | 7 | 7 |

Dataset of Kessler (2001)

# Testing the Impact of Sample Size on Cognate Detection

# Gold Standard

## IDS-Testset

- 4 languages (German, English, Dutch, French)
- 550 items (glosses)
- translations taken from the IDS (Key & Comrie 2009)
- orthographic entries converted into IPA transcriptions
- cognate judgments follow traditional literature

# Subsets of Varying Samplesize

## Creating the Subsets

Starting from the basic dataset, subsets of the data were created by

- randomly deleting 5, 10, 15, etc. items from the original dataset, and
- taking 5 different samples for each distinct number of deletions.

This process yielded 550 datasets, covering the whole range of possible sample sizes between 5 and 550 in steps of 5.

# Automatic Cognate Detection

## Methods for Cognate Detection

- Normalized Edit Distance (NED)
- Turchin et al. (2010, Turchin)
- SCA Distance (List 2012b)
- LexStat (List 2012a)

## Implementation

All methods are implemented as part of LingPy-1.0 (see `http://lingpy.org`), a Python library for quantitative tasks in historical linguistics.

# Evaluation Measures

## B-Cubed Precision and Recall (Amigó et al. 2009)

Given a test (result of an analysis) and a reference (the gold standard),

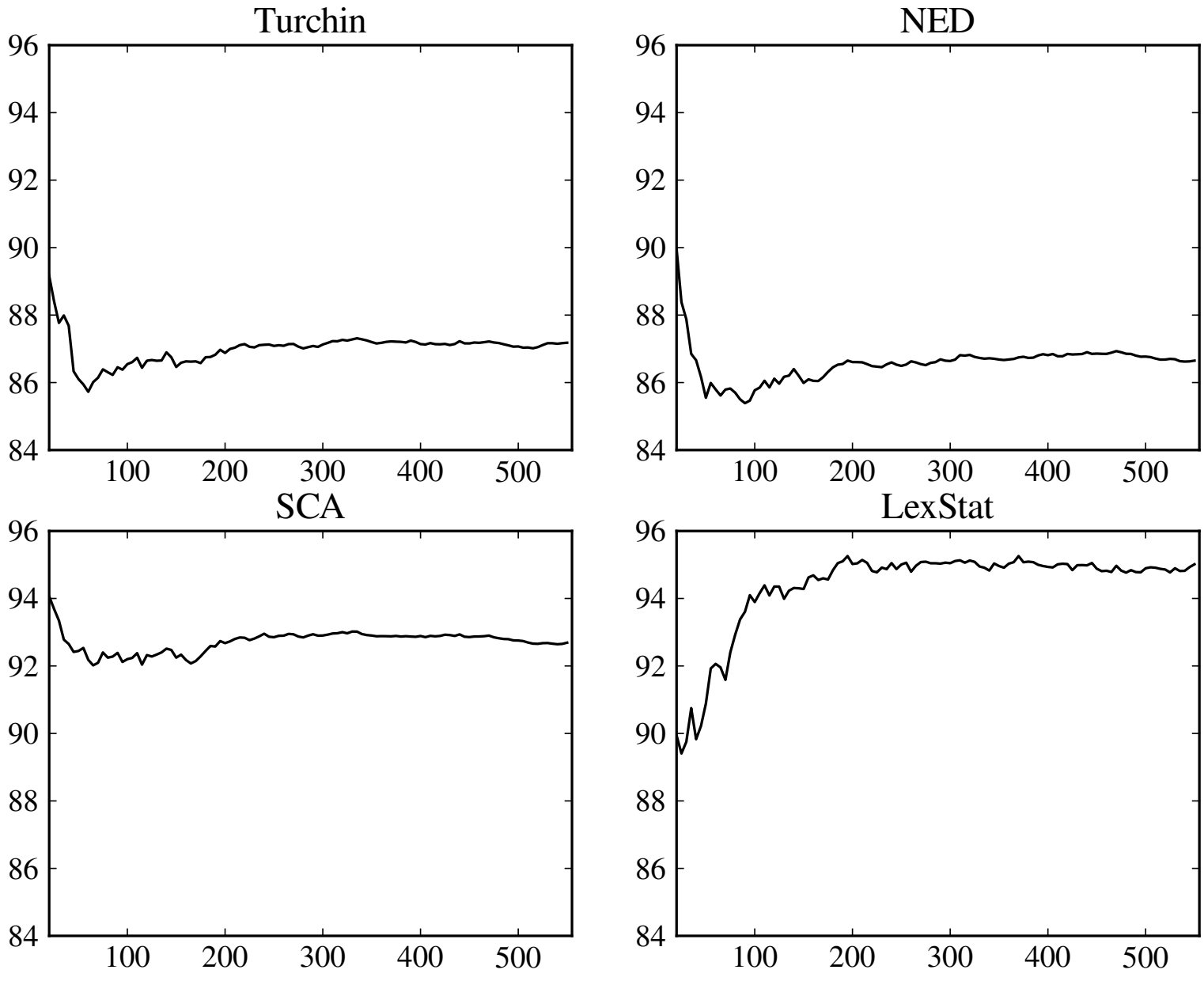
- precision is the proportion of items in the test that also occur in the reference, and
- recall is the proportion of items in the reference that also occur in the test.

Low precision is equivalent to high rates of false positives, low recall is equivalent to high rates of false negatives (missed cognates).

# Results

# Results

| Items | B-Cubed Recall | | | |
|---|---|---|---|---|
| | Turchin | NED | SCA | LexStat |
| 50 | 86.10 | 85.55 | 92.44 | 90.88 |
| 100 | 86.55 | 85.77 | 92.20 | 93.89 |
| 200 | 86.88 | 86.61 | 92.68 | 95.02 |
| 300 | 87.13 | 86.64 | 92.90 | 95.05 |
| 400 | 87.14 | 86.81 | 92.89 | 94.94 |
| 500 | 87.07 | 86.77 | 92.75 | 94.90 |


Turchin
NED
SCA
LexStat
96
94
92
90
88
86
84
100
200
300
400
500

# Discussion

# Discussion

## Are 200 words enough?

Although

- the representativity of the data is limited, and
- the number of languages investigated is small,

the test shows that

- sample size has a definite impact on the results of language-specific methods, and
- using 200 words is surely better than using 100 words.

| | | | | | | |
|---|---|---|---|---|---|---|
| Sanscruta | *sarpá-* | s | a | r | p | a |
| Italienisch | *serpe* | s | ɛ | r | p | ə |
| Sanscruta | *devá-* | d | e | v | a | |
| Italienisch | *Dio* | d | i | - | o | |
| Sanscruta | *saptá-* | s | a | p | t | a |
| Italienisch | *sette* | s | ɛ | - | tː | ə |

# Спасибо за Ваше Внимание!