

Vocabulary Testing: some methodological considerations

Norman D. Verhelst

National Institute for Educational Measurement (Cito)

Arnhem – The Netherlands

Fifth EALTA conference

Athens

May 9-11 2008

General Context

- Important domain
 - Interest sui generis
 - Compare Dialang
- Computerized testing
 - ‘Electronize’ P&P test
 - New features
- Computer adaptive testing (CAT)
 - Using IRT
 - Other methodology

Fixed vs. Random

- In Classical Test Theory, items are treated statistically as fixed effects
 - Standardized tests (e.g. 80 vocab. items)
 - Reliability (correlation between two indep. administrations of the **same** test)
- Generalizability theory (Section E):
 - Items are considered as a **random sample** from a **universe** of items
 - How is the universe defined?
 - How is the sampling done?

IRT: items are fixed

- Estimating the difficulty parameter of items
 - ‘cow’ is easier than ‘horse’?
 - Costs a lot of money
 - Assumes a linear ordering of all items common to all members of the target population
- Can we get rid of all this?
 - And at the same time preserving something useful? (e.g. CAT)

A simple answer

- Let us define a **universe** of items (N)
- The true score of a student is the proportion of items 'known' (π)
- The test consists of a **random** sample of n items from the universe
- The observed score is the **proportion** of correct answers (p)

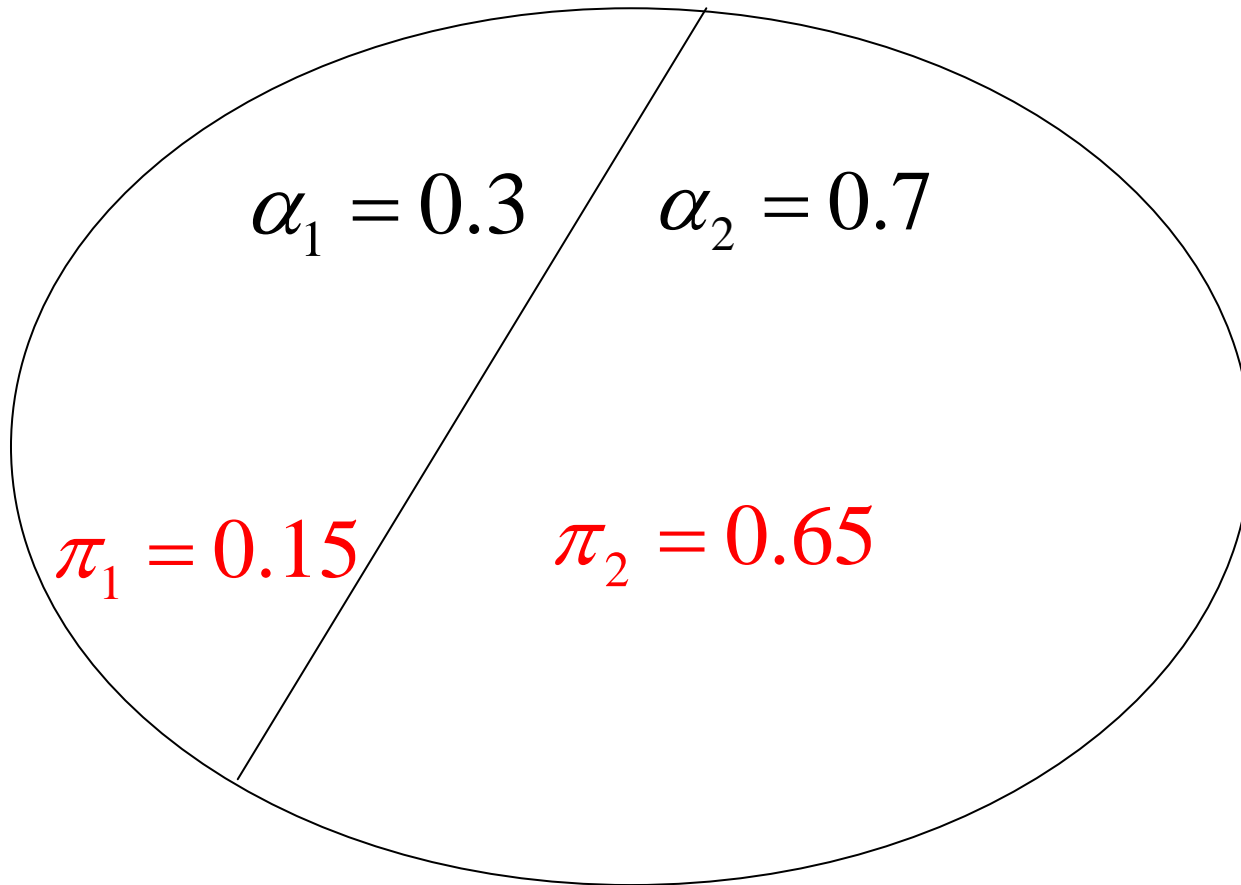
Standard error

$$SE(p) = \sqrt{\frac{\pi(1-\pi)}{n}} \approx \sqrt{\frac{p(1-p)}{n}}$$

Example: $n = 81$; $p = 0.53$

$$SE(p) \approx \sqrt{\frac{0.53 \times 0.47}{81}} = 0.055$$

Stratification



$$\pi = \alpha_1 \pi_1 + \alpha_2 \pi_2 = 0.3 \times 0.15 + 0.7 \times 0.65 = 0.50$$

Accuracy and stratification

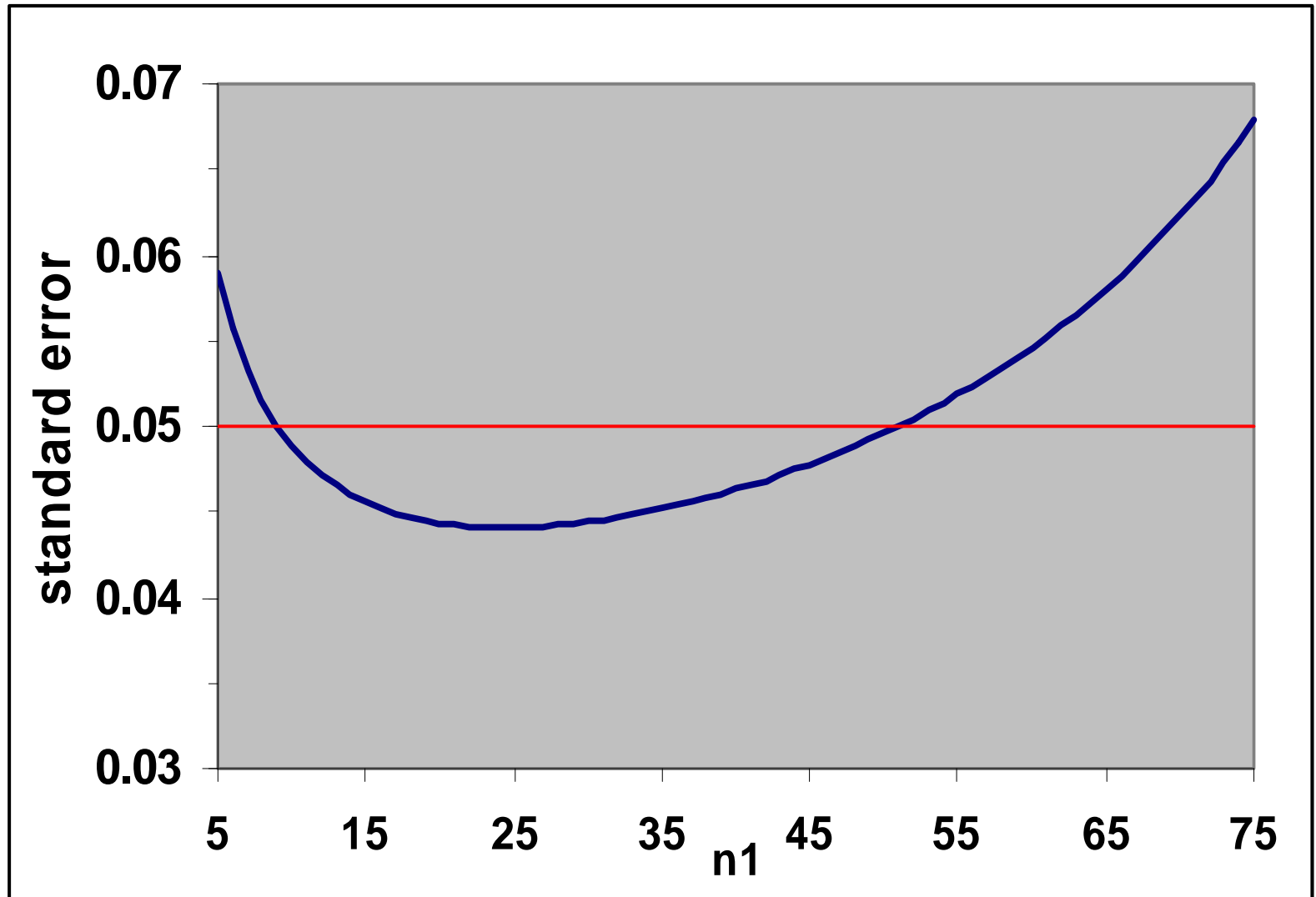
Suppose n is given (100, say)

n_1 items from stratum 1

$n_2 = n - n_1$ items from stratum 2

$$p = \alpha_1 p_1 + \alpha_2 p_2$$

$$SE(p) = \sqrt{\frac{\alpha_1^2 \pi_1 (1 - \pi_1)}{n_1} + \frac{\alpha_2^2 \pi_2 (1 - \pi_2)}{n - n_1}}$$



Standard error reaches its minimum at $n_1 \approx 24$

Decision rule

$$A_1 = \alpha_1 \sqrt{\pi_1(1 - \pi_1)} = 0.3 \times \sqrt{0.15 \times 0.85} = 0.107$$

$$A_2 = \alpha_2 \sqrt{\pi_2(1 - \pi_2)} = 0.7 \times \sqrt{0.65 \times 0.35} = 0.334$$

$$\frac{A_1}{A_1 + A_2} = \frac{0.107}{0.107 + 0.334} = 0.243$$

To reach minimal standard error **24.3%**
of the items are sampled from stratum 1

Efficiency

For this student, we get the same accuracy

- 100 items with stratified sampling
- 129 items with simple random sampling

	Str. 1	Str.2
α	0.3	0.7
n	16	44
#correct	3	29
p	0.188	0.659
A	0.117	0.332
$A / \Sigma A$	0.261	0.739

item 61

Str. 1: $17 / 61 = 0.279$

Str. 2: $16 / 61 = 0.262$



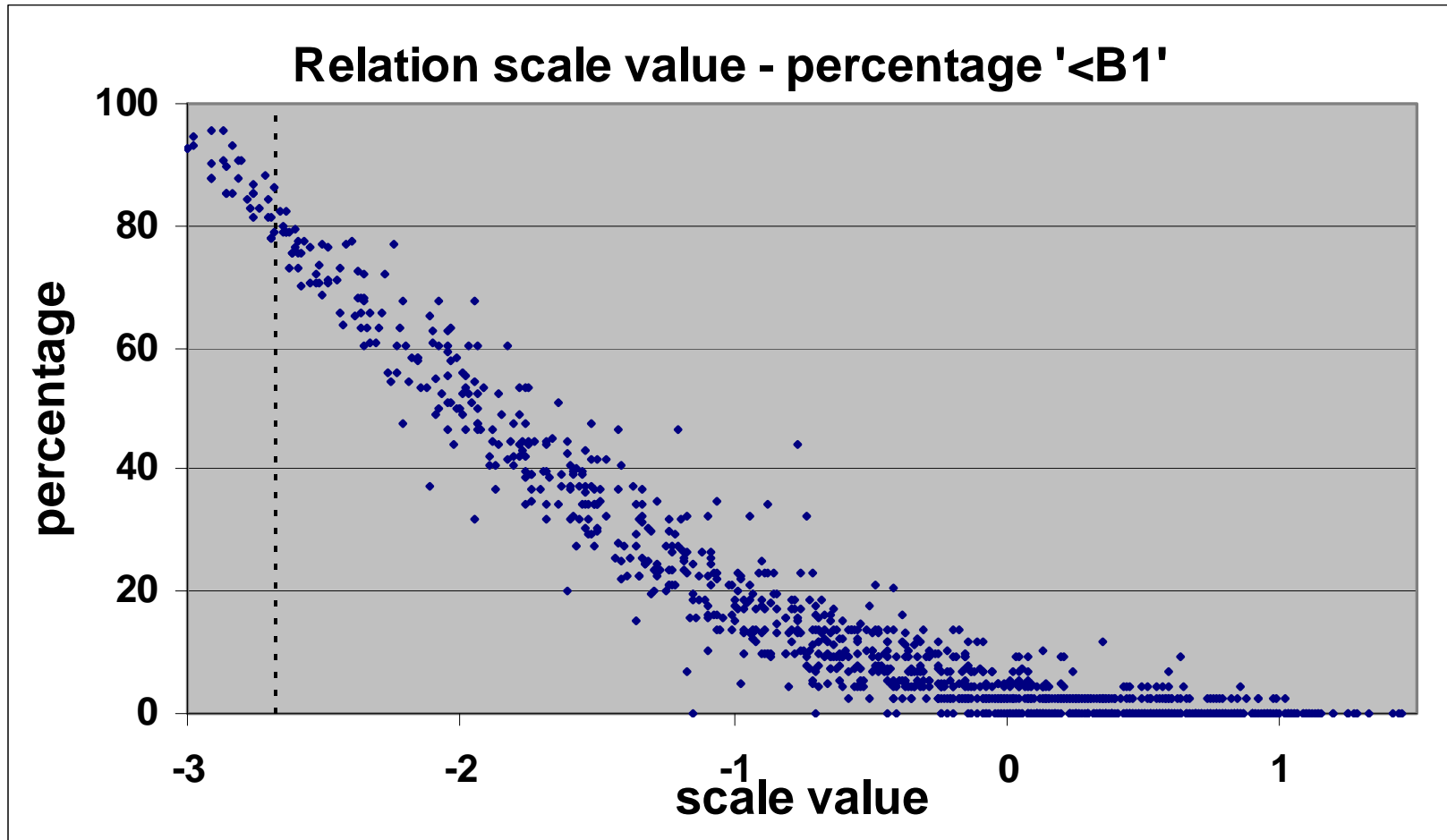
Building a Universe

- Define a population of texts and ‘words’
 - Textbooks
 - Books and magazines (fiction/non-fiction)
 - Word frequency lists
 - Language variation
- Define the ‘depth’ of vocabulary items
 - Denotation
 - Connotation
 - Register
 - Semantic network
 - Idiom
 - Morphology
 - Non-language words

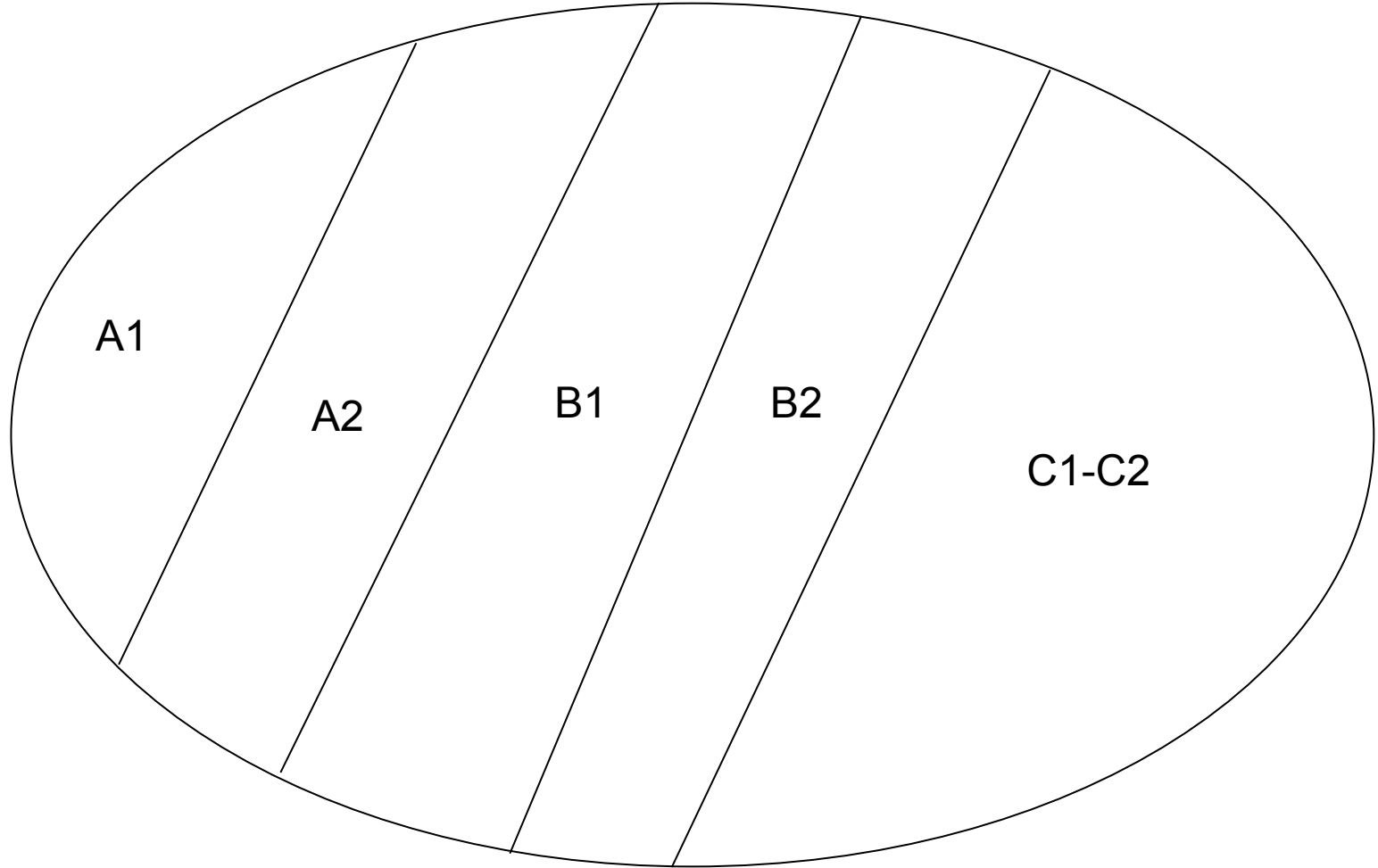
Stratification of the universe: an example

- (Dutch) vocabulary in Dutch primary education
- 2400 words and expressions
- ‘passive recognition’
- Teachers judge for about 25% of the words at which grade they should be mastered (example)
- Words are scaled (homogeneity analysis)

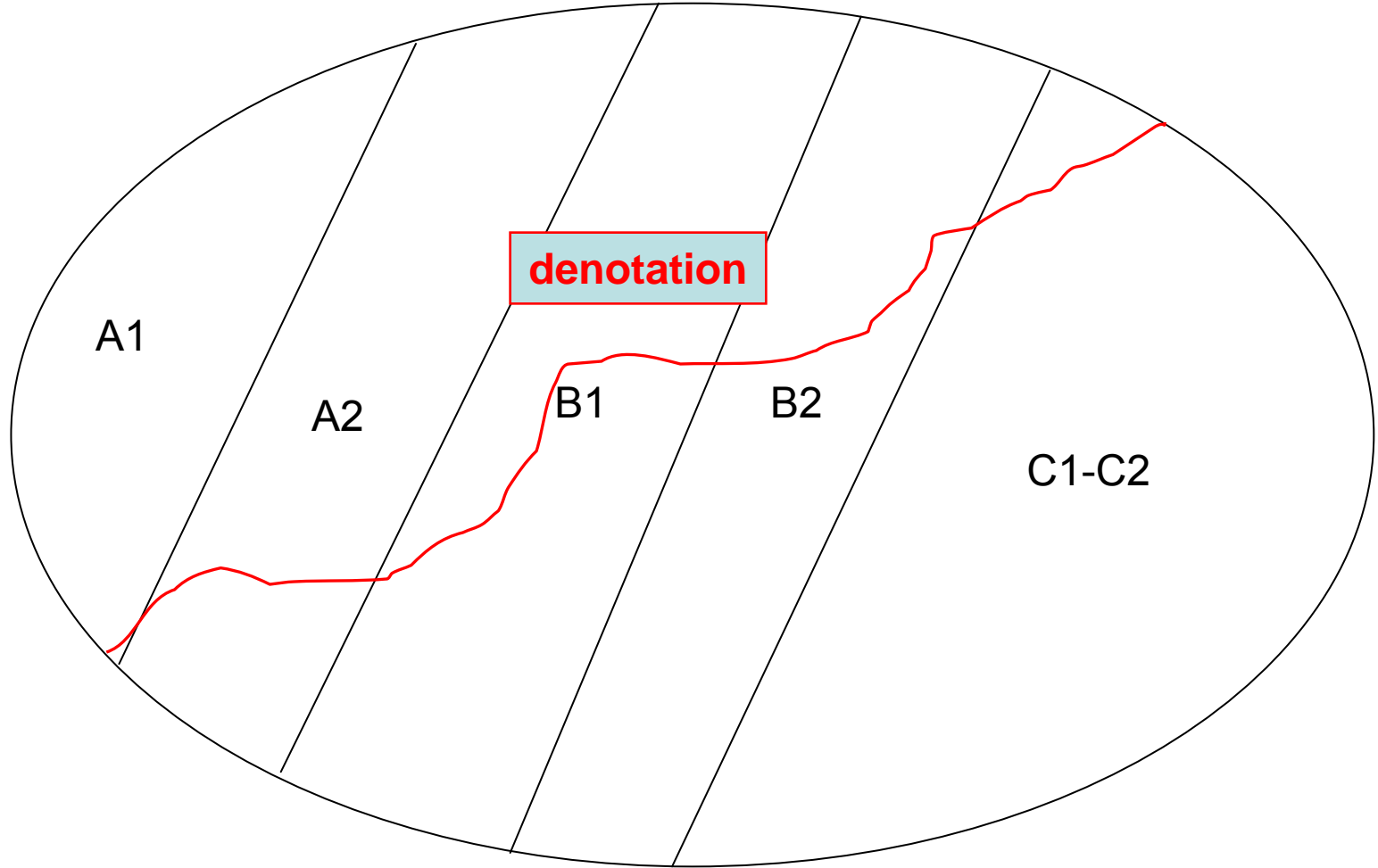
Example of results



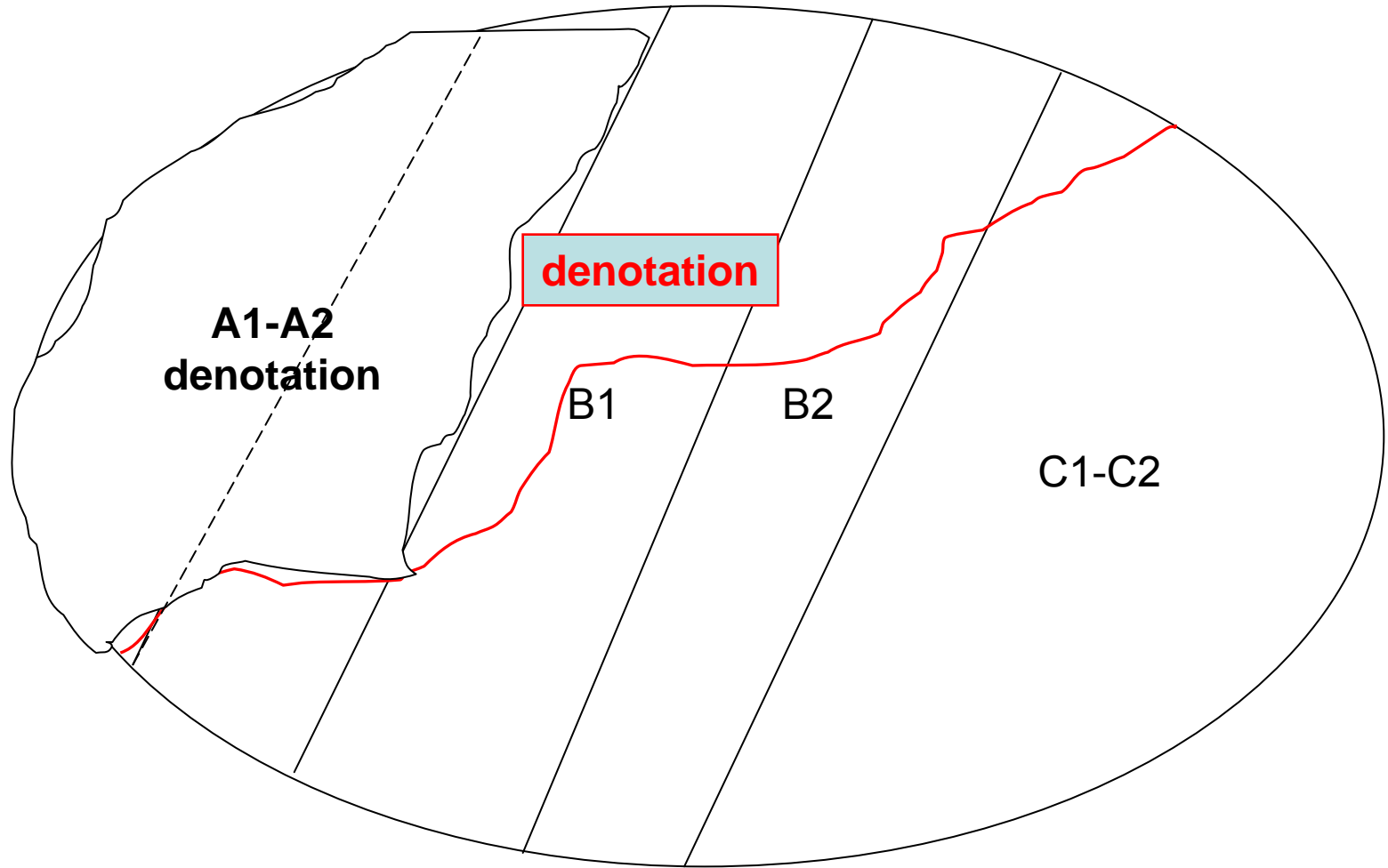
Stratification (1)



Stratification (2)



Sub-Universes



Proposal

- Building a European Vocabulary item bank
 - Start with one language
 - #lemmas approx. 8,000
 - #items approx. 15,000
- Basic stratification(s)
 - Based on expert judgment
 - Defining meaningful sub-universes
- Standard setting
- Open software
- Common data base

Proposed time line

- Year 1: making up your mind
- Year 2: forming a consortium and acquisition of funding
- Year 2-4: building the bank, testing
- Year 5: Release

Thank you