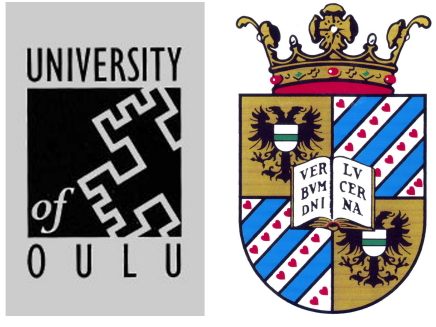
Detecting Syntactic Differences in Spontaneous Conversation

A Robust Corpus Comparison Technique

John Nerbonne, Timo Lauttamus and
Wybo Wiersma

Rijksuniversiteit Groningen & University of Oulu
j.nerbonne@rug.nl, wybo@logilogi.org, timo.lauttamus@oulu.fi,

8 July 2009
Middle and Modern English Corpus Linguistics



In this talk we present

A method for detecting syntactic differences and our findings on pausing

3 sub-questions about the method

1. What did your corpus look like ?
2. What is permutation statistics ?
3. How to apply it to syntax ?

3 sub-questions about the results

1. What general differences did you find ?
2. How much pausing is there, and by who ?
3. What does this tell about the speakers ?

Outline of the Talk

The Method

The Method

- Detect a wide range of syntax differences, and these as
  - significant differences
  - aggregate differences
  - relative differences
- This would enable measuring the syntax part of total impact:

"No easy way of measuring or characterizing the total impact of one language on another in the speech of bilinguals has been, or probably can be devised. The only possible procedure is to describe the various forms of interference and to tabulate their frequency."

The Method

- Detect a wide range of syntax differences, and these as
  - significant differences
  - aggregate differences
  - relative differences
- This would enable measuring the syntax part of total impact:

"No easy way of measuring or characterizing the total impact of one language on another in the speech of bilinguals has been, or probably can be devised. The only possible procedure is to describe the various forms of interference and to tabulate their frequency."

- Weinreich, *Languages in Contact*

The Method

We also want to do this automatically and computationally
In order to be able to:

- Mine for differences in syntax between
  - learners versus native speakers
  - speakers of different dialects
  - writers from different discourses
- Test dialectological and other linguistic hypotheses
- Note over- and under-use instead of right / wrong

The Method

We did it in four steps:

1. Tag 2 or more collections of comparable material (using an automatic POS-tagger)
2. Take n-grams (2 - 5 grams) of POS-tags
3. Statistically compare their frequencies
4. Sort the significant POS-n-grams by extent of difference

Aarts J. and Granger S. did this without the statistics in: 'Tag sequences in learner corpora: a key to interlanguage grammar and discourse' (1998)

Our Corpus

Origins:

- 20,000 Finns immigrated to Australia
- Working class background, limited education
- 25-40 Years upon arrival

Corpus collected 1995-1998 by Greg Watson:

- of the university of Joensuu, Finland
- two age groups; adults and juveniles
- 350.000 words, 305.000 words free conversation

Our Corpus

Adults:

- over 18 years at arrival, on average 30
- on average 58 at time of interview
- 60 interviews, 65 - 70 min each (221.000 words)

Juveniles:

- under 16 years at arrival, on average 6
- on average 36 at time of interview
- 30 interviews, 65 - 70 min each (84.000 words)

Our Corpus

In preparation we Part of Speech-tagged it with:

- Trigrams 'n 'Tags (TnT) Statistical POS Tagger
- made by Thorsten Brants (Universitt des Saarlandes)

It achieves an accuracy of:

- 96.7% on the Penn Treebank
- 85.1% - 90.5% on our spoken material

Accuracy is of course worse for 3-grams:

- 2-grams 74%, 3-grams 65%, 4-grams 58% ...

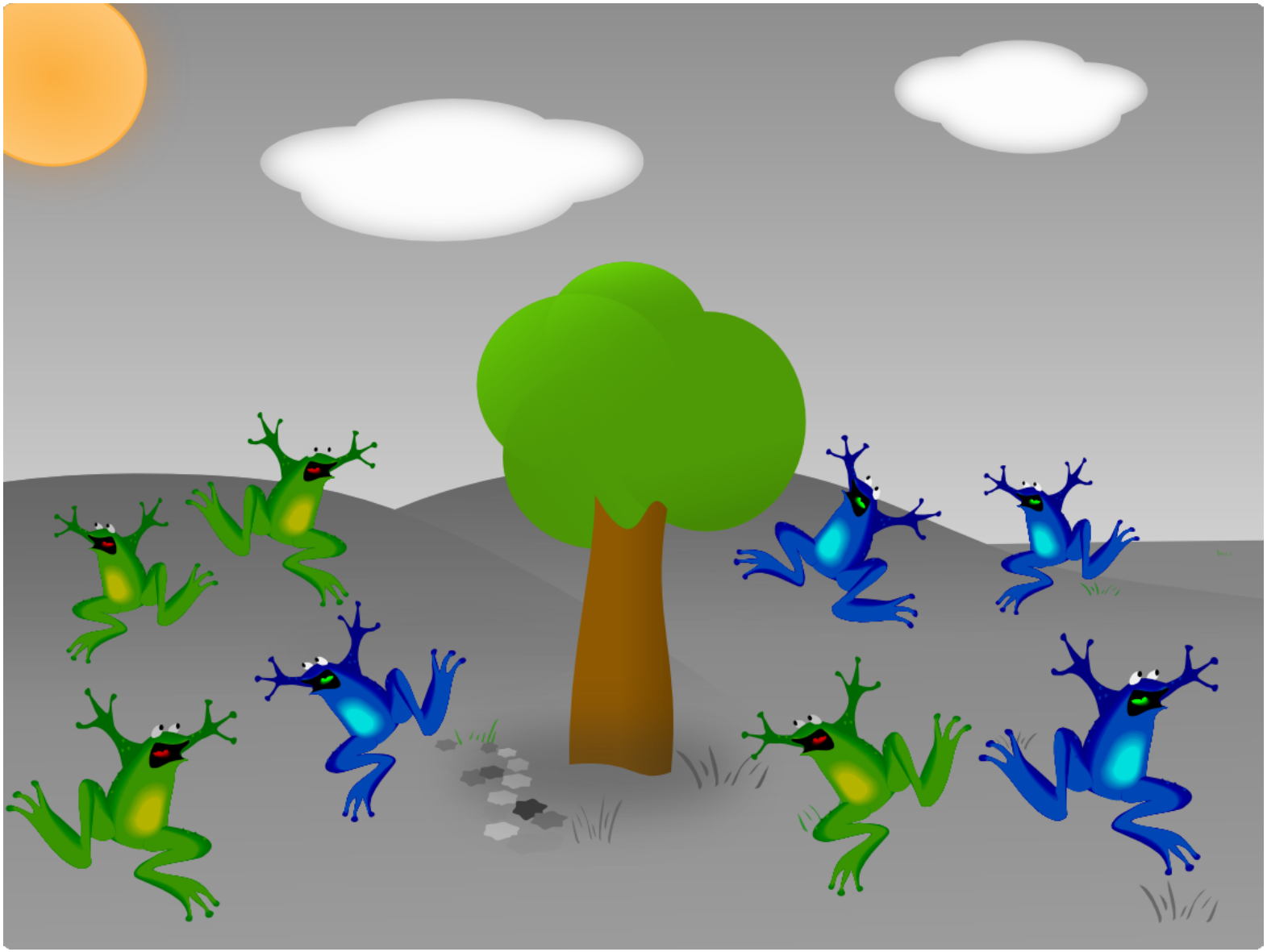
Permutation Statistics

It is different from parametric (normal) statistics:

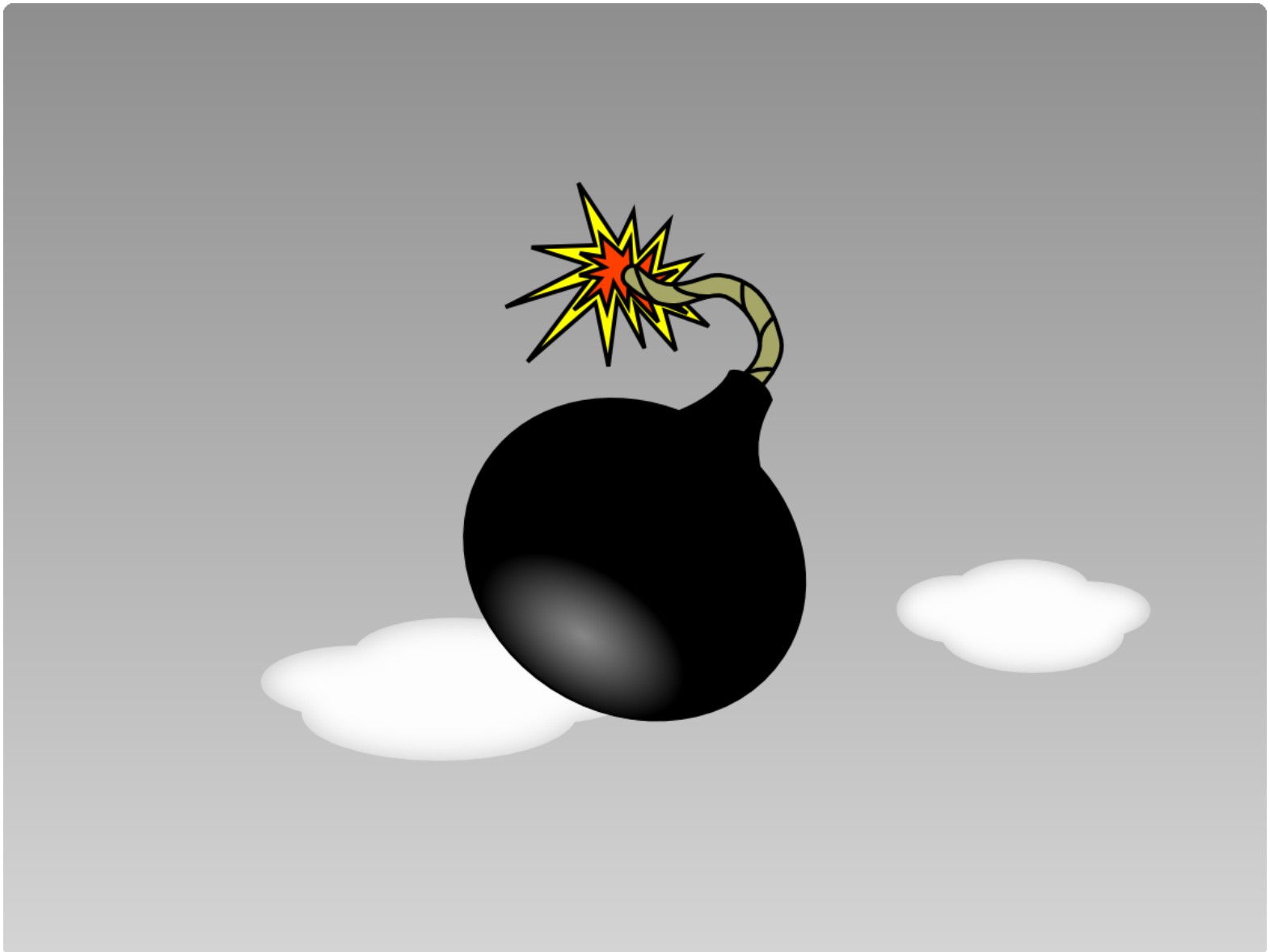
- It is about the data, not about the population
  - no need for normality
  - no need for homoscedasticity (eq distrib variances)
  - no absolute need for random sampling
- Still, important for permutation statistics are
  - random assignment and independence of observations
  - in practice no problems for linguistic/dialect data

As a statistical method it is very suitable for linguistics

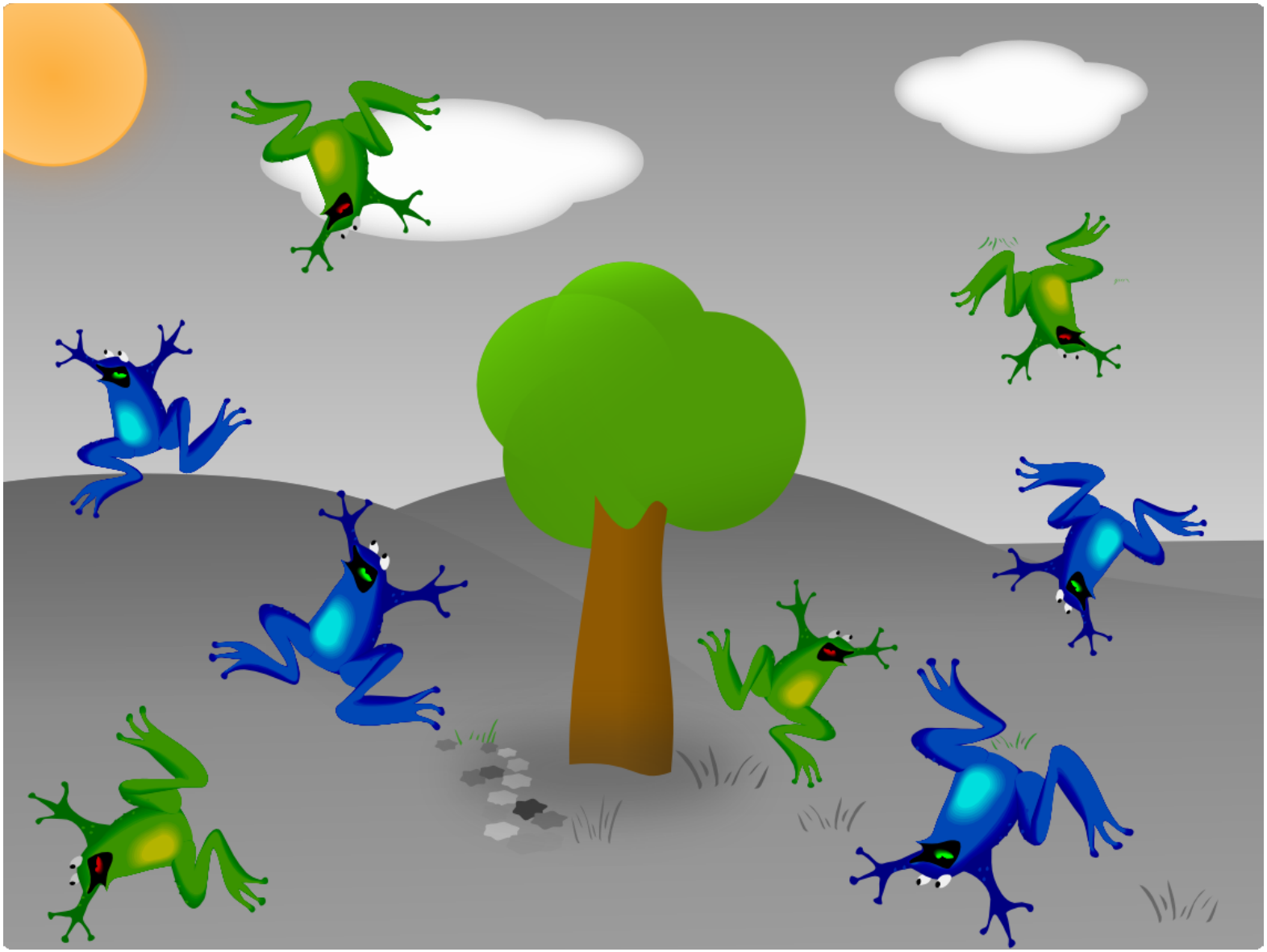
Permutation Statistics



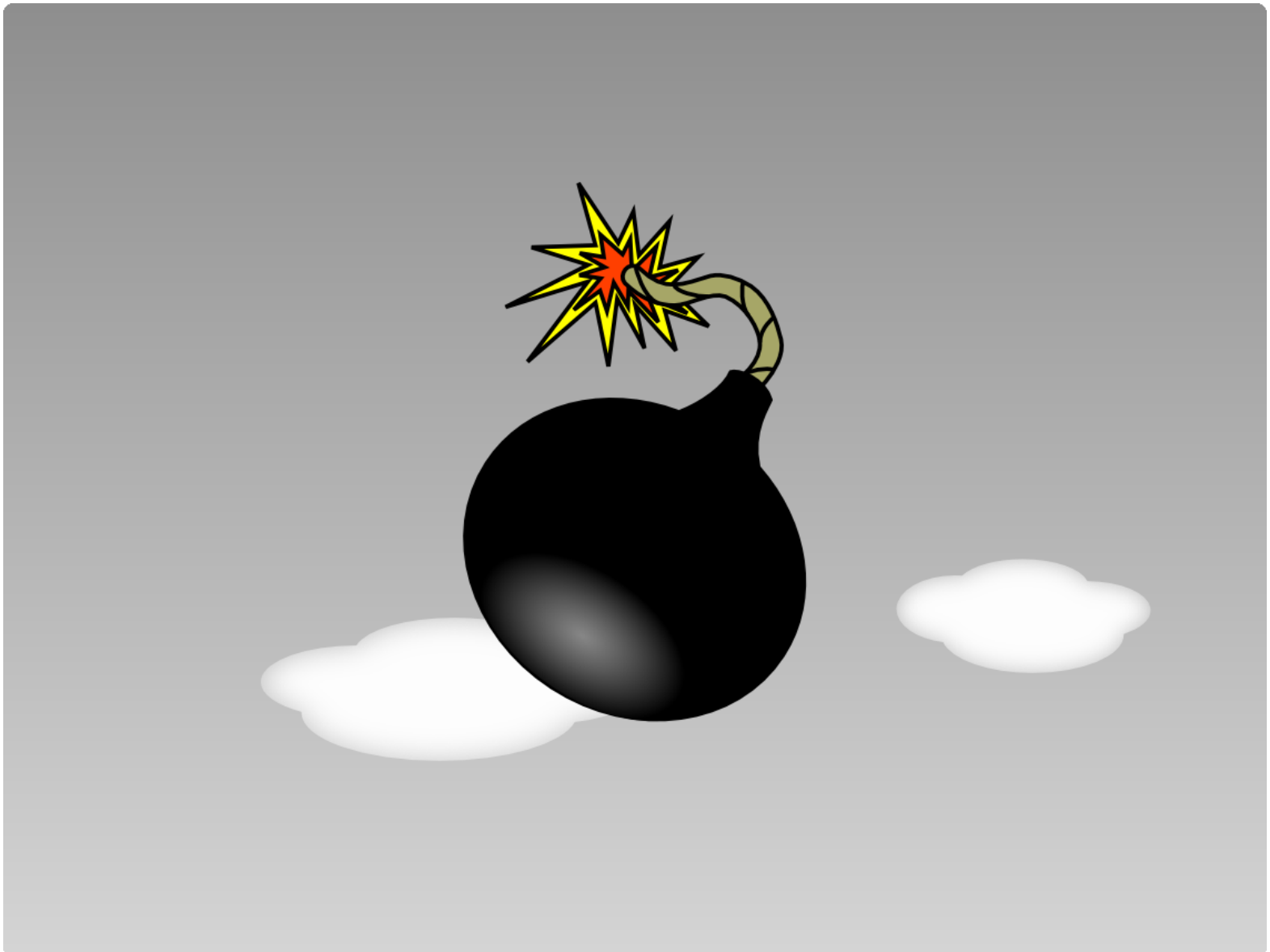
Permutation Statistics



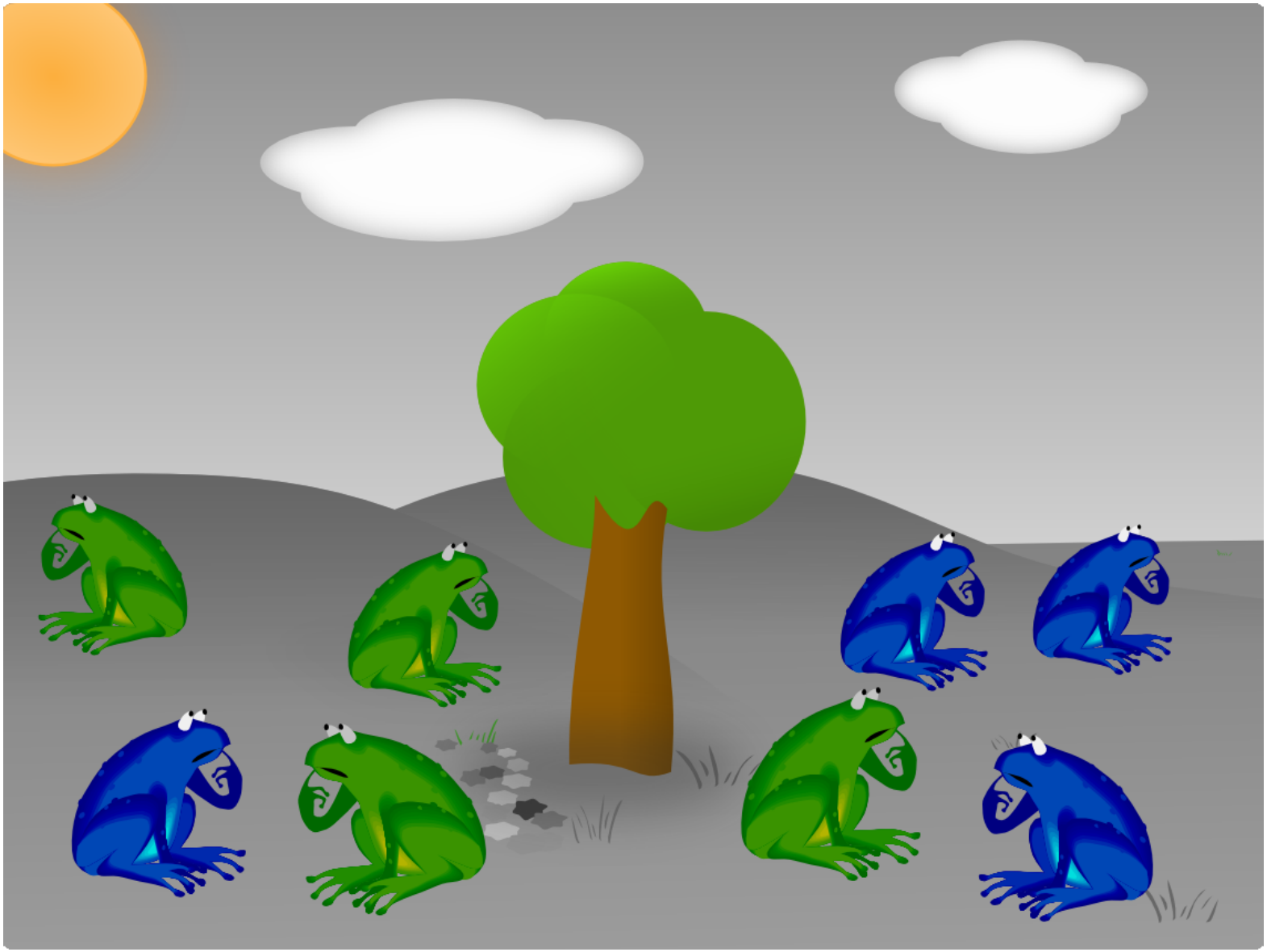
Permutation Statistics



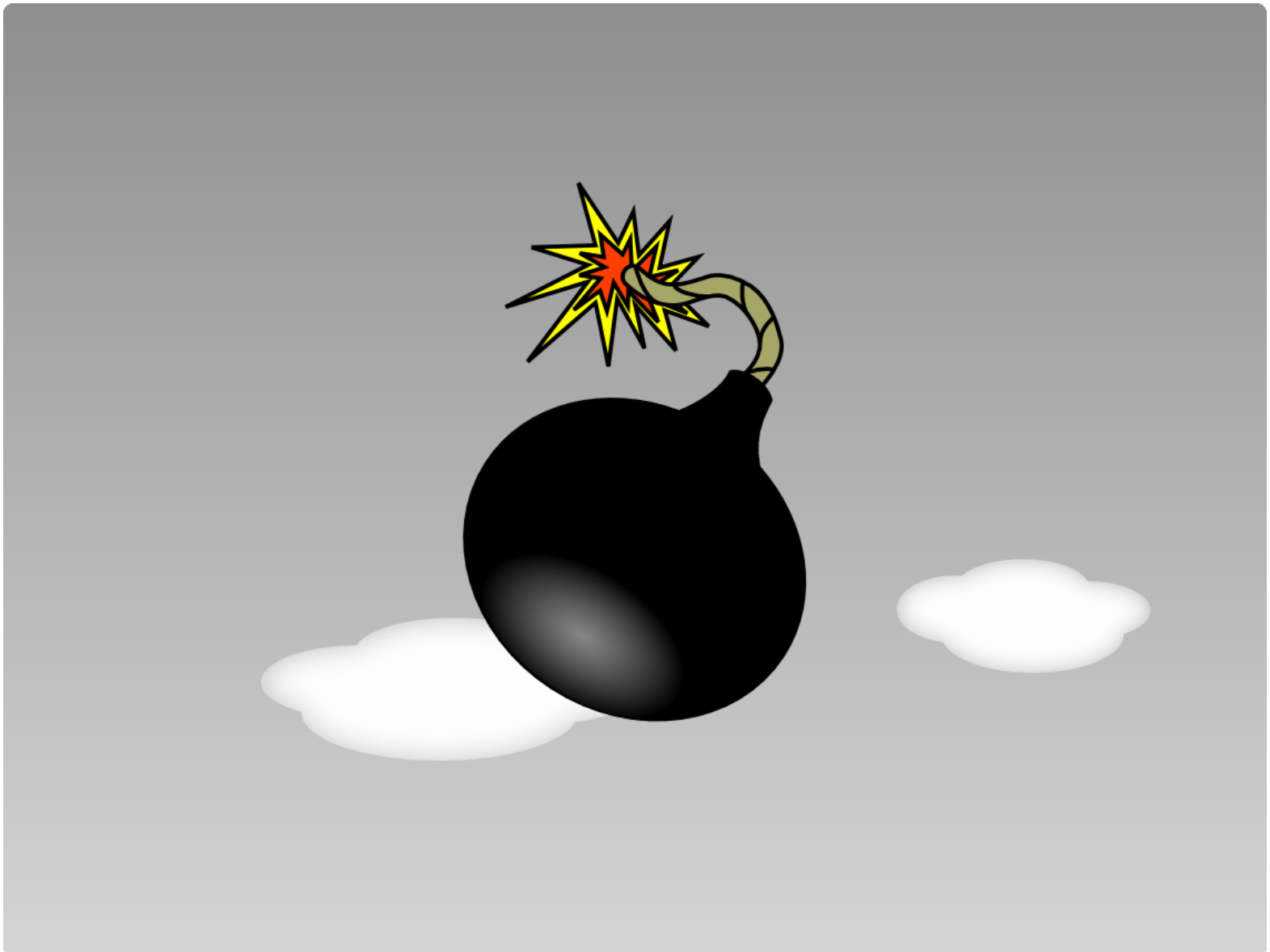
Permutation Statistics



Permutation Statistics

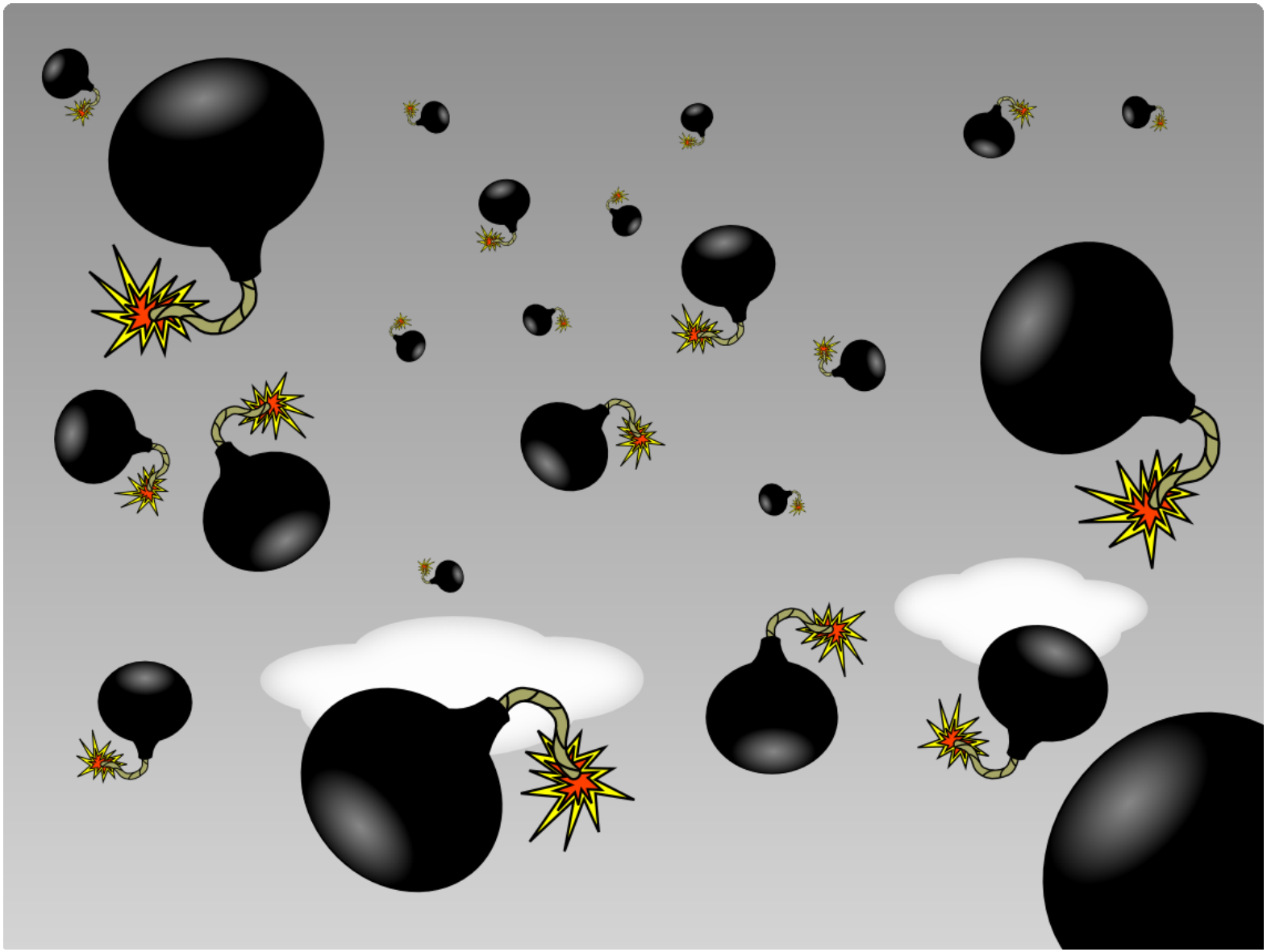


Permutation Statistics

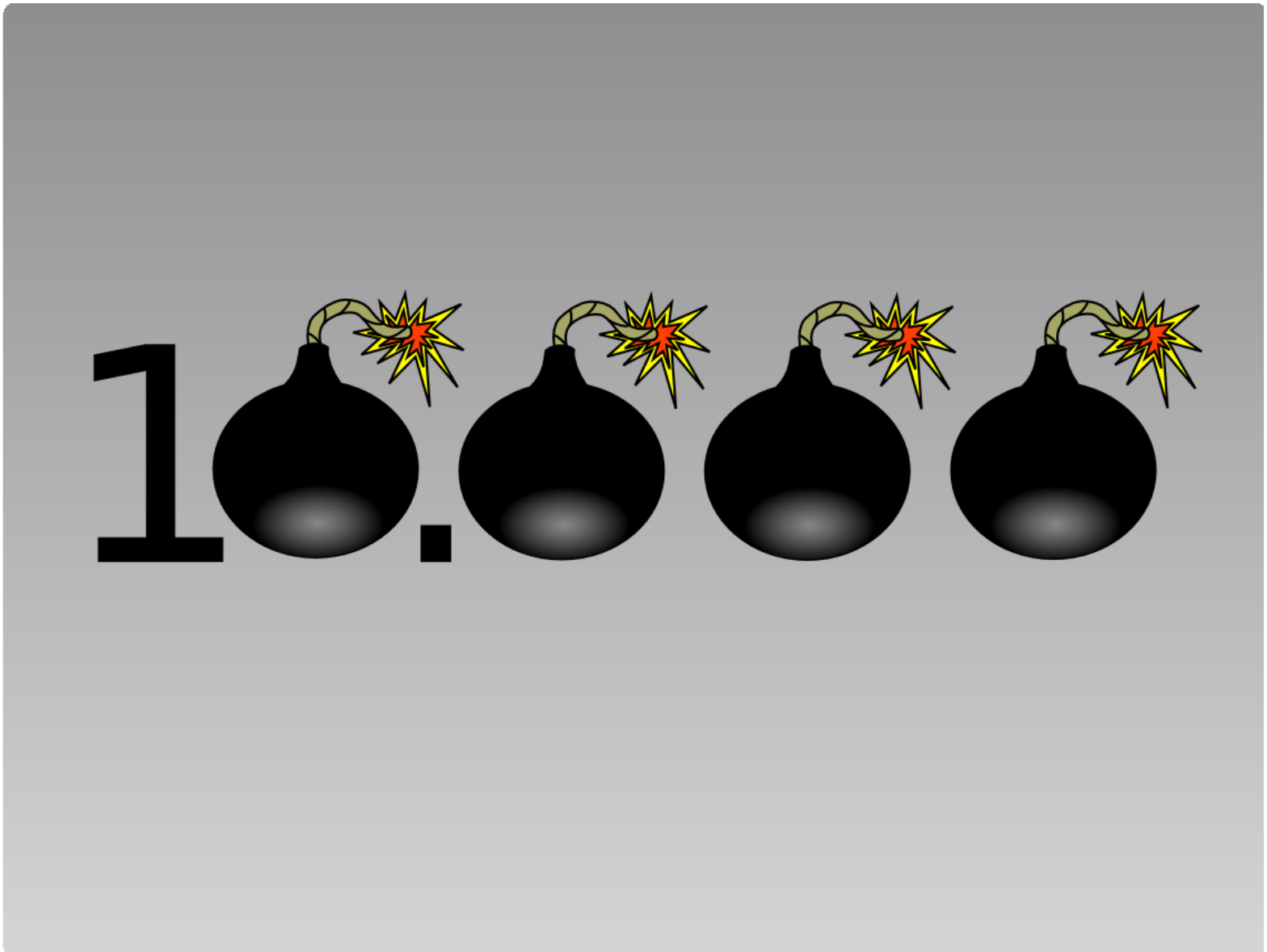


Permutation Statistics

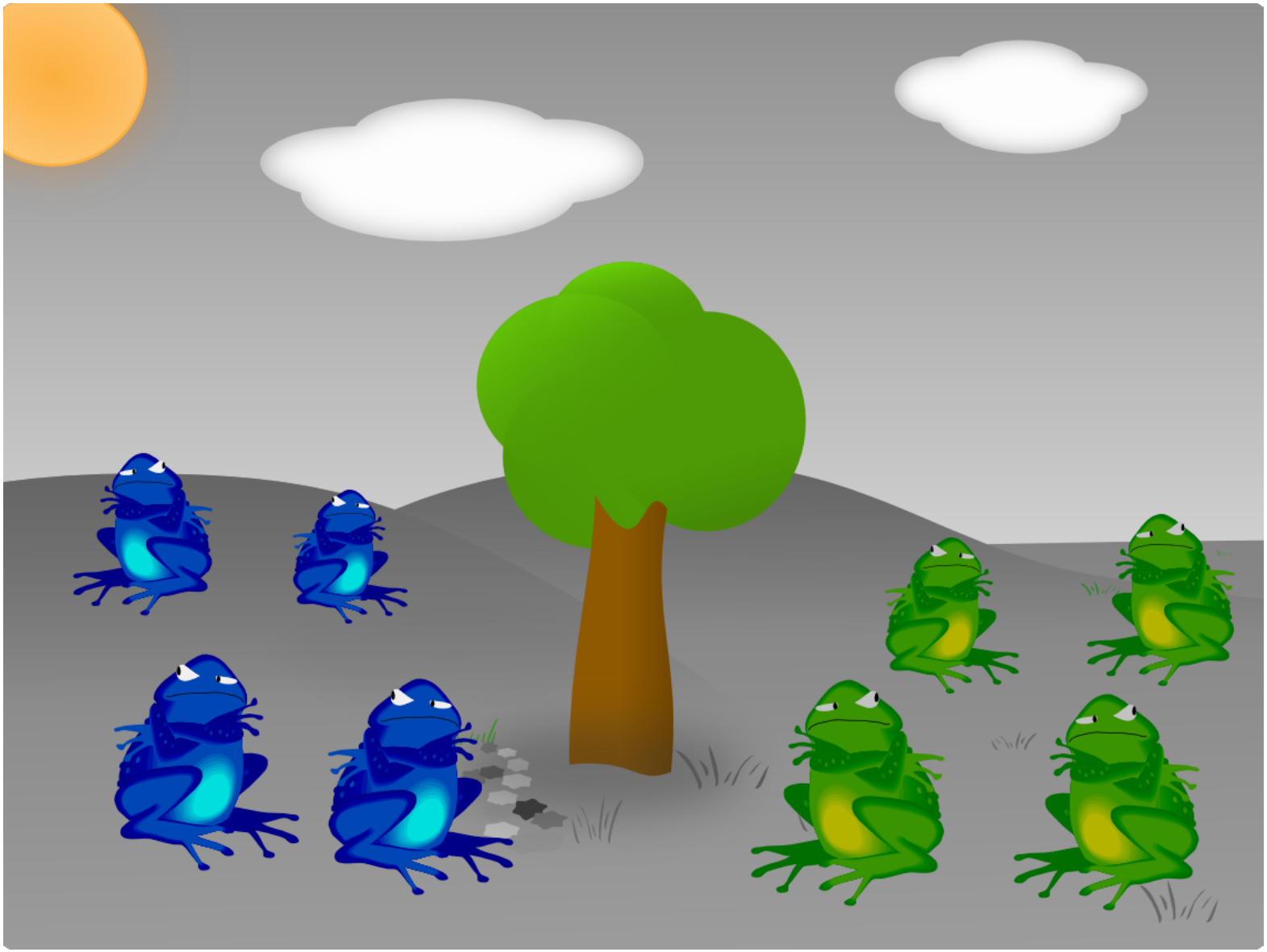
Permutation Statistics



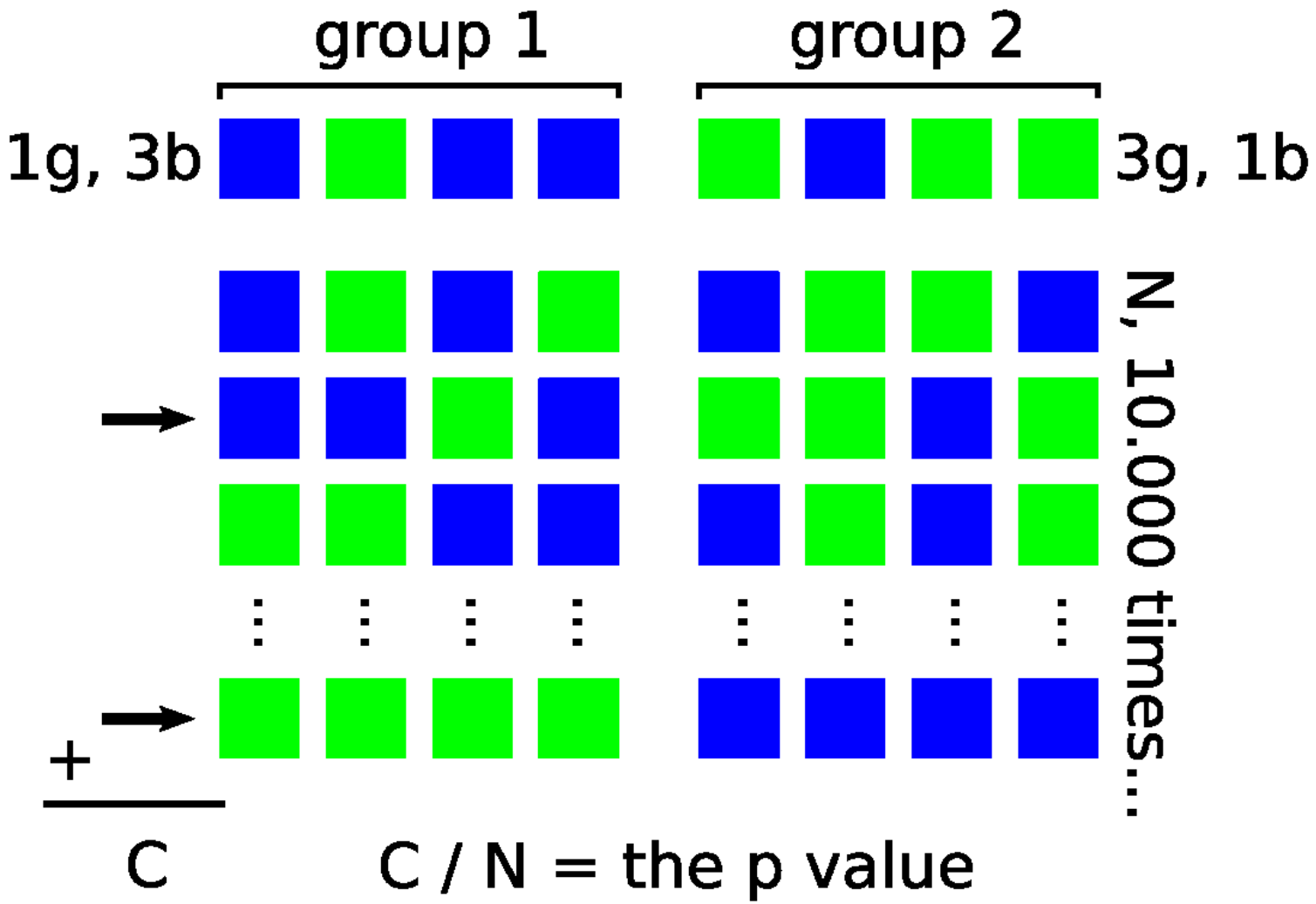
Permutation Statistics



Permutation Statistics



Permutation Statistics



Applying it to Syntax

One firstly needs something to permutate:

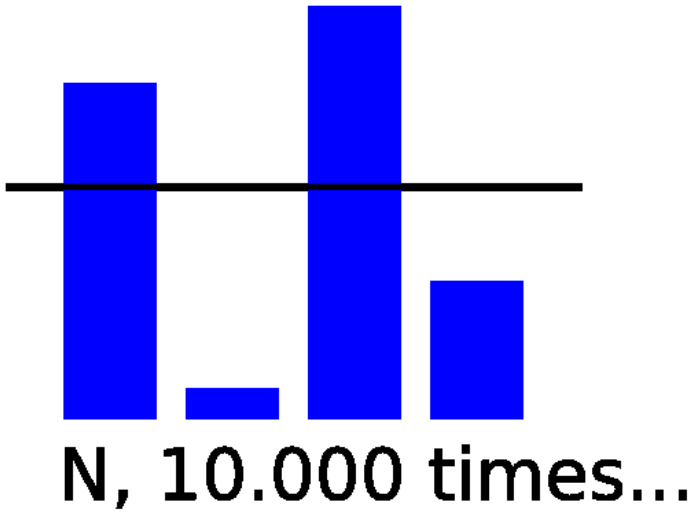
- We permutated interviewees
  - more conservative than 3-grams
  - and also easier than sentences (did this earlier)
- For each interview
  - we took 3-grams (N-grams too) of POS-tags
  - we then calculated the 3-gram-promillages for all 3-grams (occurrence of 3-gram type per 1000 3-gram tokens)
- These 3-gram-promillage-vectors were then used
  - summed per group after each permutation

Applying it to Syntax

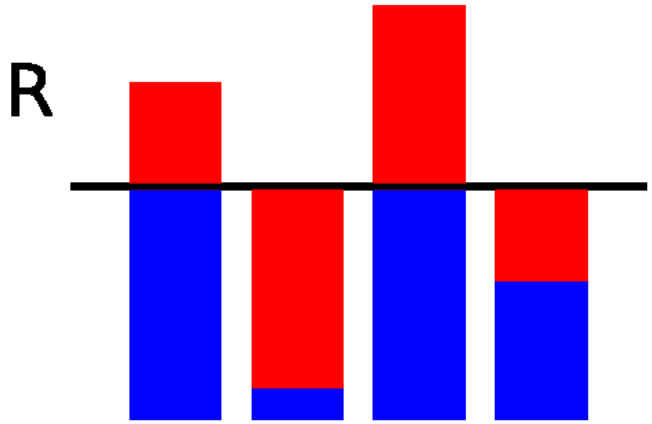
Secondly one needs something to measure extremity:

- Both
  - for the whole group
  - and for each individual POS-3-gram
- We used r-square and summed r-square
  - we also tried cosine and summed r
- R-square is the square of the difference (r)
  - for a POS-3-gram-promillage between the 2 groups
- Summed r-square is the sum of r-square for all 3-grams

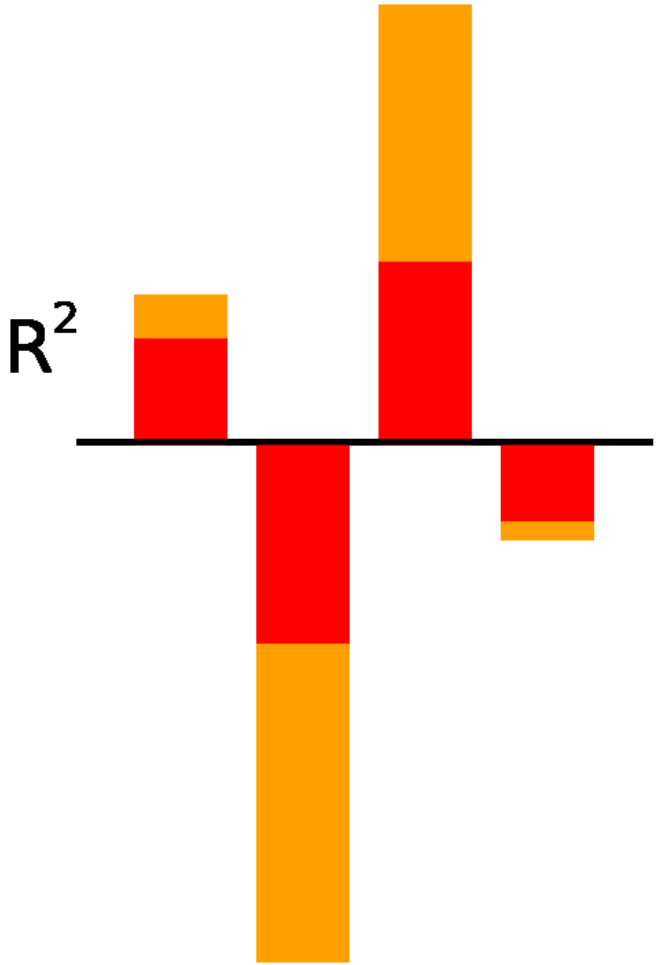
R-square



N, 10.000 times...



N, 10.000 times...



N, 10.000 times...

Applying it to Syntax

Thirdly one needs to apply normalizations for:

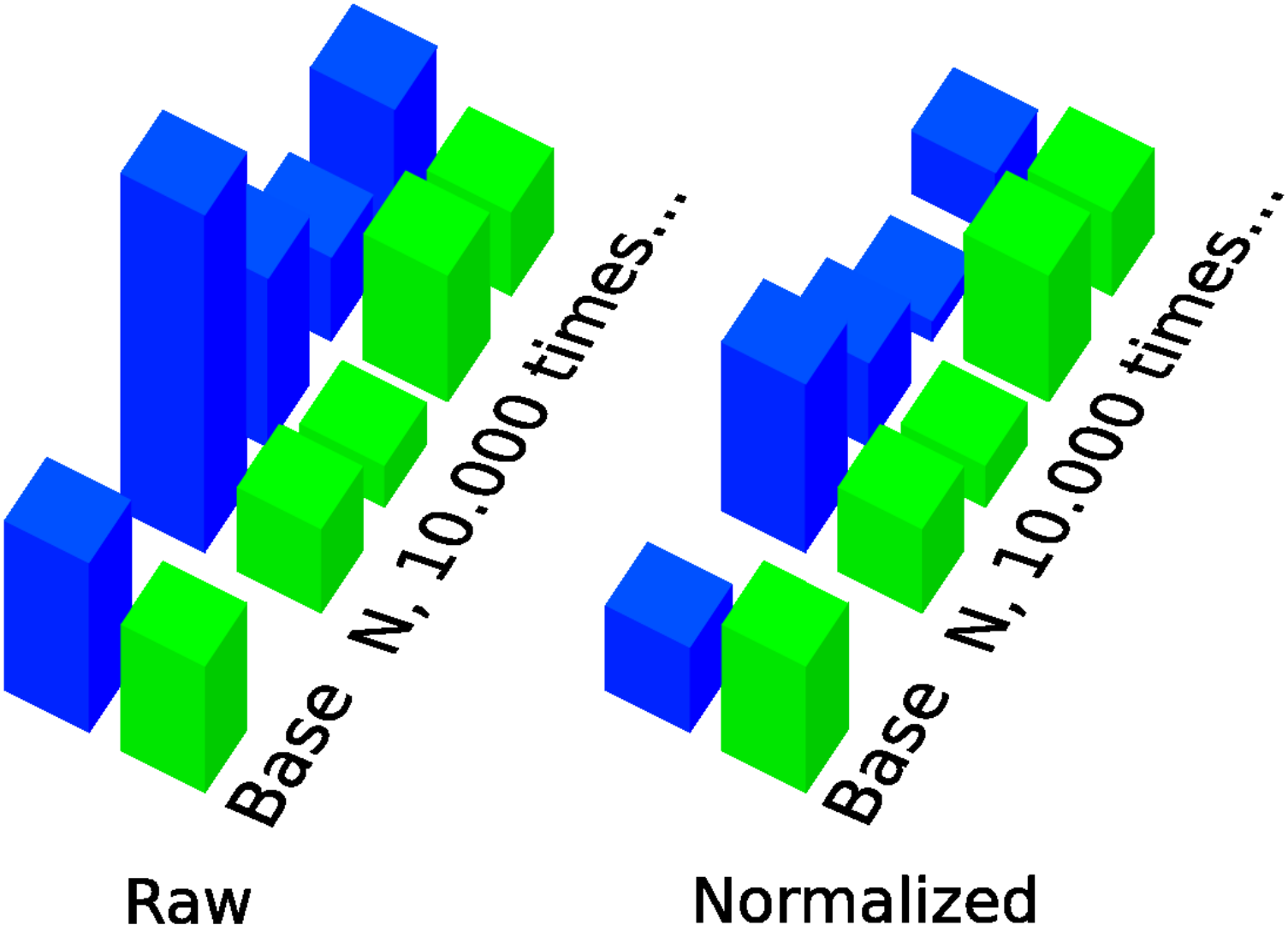
- Text-size per subject (for each subject)
  - divide by sum of the subjects' 3-grams (the promillages)
  - to eliminate differences in text-size between authors
- Frequencies of 3-gram types (for each 3-gram type)
  - divide by the corpus-wide total of the 3-gram-type
  - to eliminate differences in frequencies (optional)
- Group-size (for both groups, across permutations)
  - divide by the average fequency of 3-grams in the group
  - to correctly detect over- and under-use

Applying it to Syntax

Normalisations are needed to

- Prevent false significance
  - arising from differences in text- and group-sizes
- Increase the weight of less frequent 3-grams
  - on the level of the group
  - (as said this is optional)
- And it allows one to sort 3-grams based on
  - whether they are more or less typical for each group
  - relative to group-size

Normalizing for Frequency



Results and Analysis

Results and Analysis

- The following slides summarise some of the material in Lauttamus, Nerbonne, and Wiersma (2007, 2009)
- The evidence based on the data of the two groups shows that there are statistically significant syntactic differences between the adult and juvenile groups
- We argue that some of the significant differences found in the data can be ascribed to the lower level of language proficiency of the adults

General Differences

Some of the syntactic differences found in the data can be described in most general terms as follows (all for the adults group):

1. Overuse of hesitation phenomena
2. Overuse of parataxis
3. Underuse of contracted forms
4. Reduced repertoire of discourse markers
5. Avoidance of complex verbal structures
6. Avoidance of prepositional and phrasal verbs
7. Underuse of the existential there

General Differences

- The adults demonstrate features of disfluent speech
  - such as (filled) pauses, repeats, false starts, incomplete or false syntactic structures, arising from difficulties in speech processing, and particularly in lexical access
- We argue that the statistical evidence obtained from our data reflects syntactic distance between the two varieties of L2 English
- And, consequently, aggregate effects of the differences in the two groups English proficiency
- We will now look further into pausing

Pausing

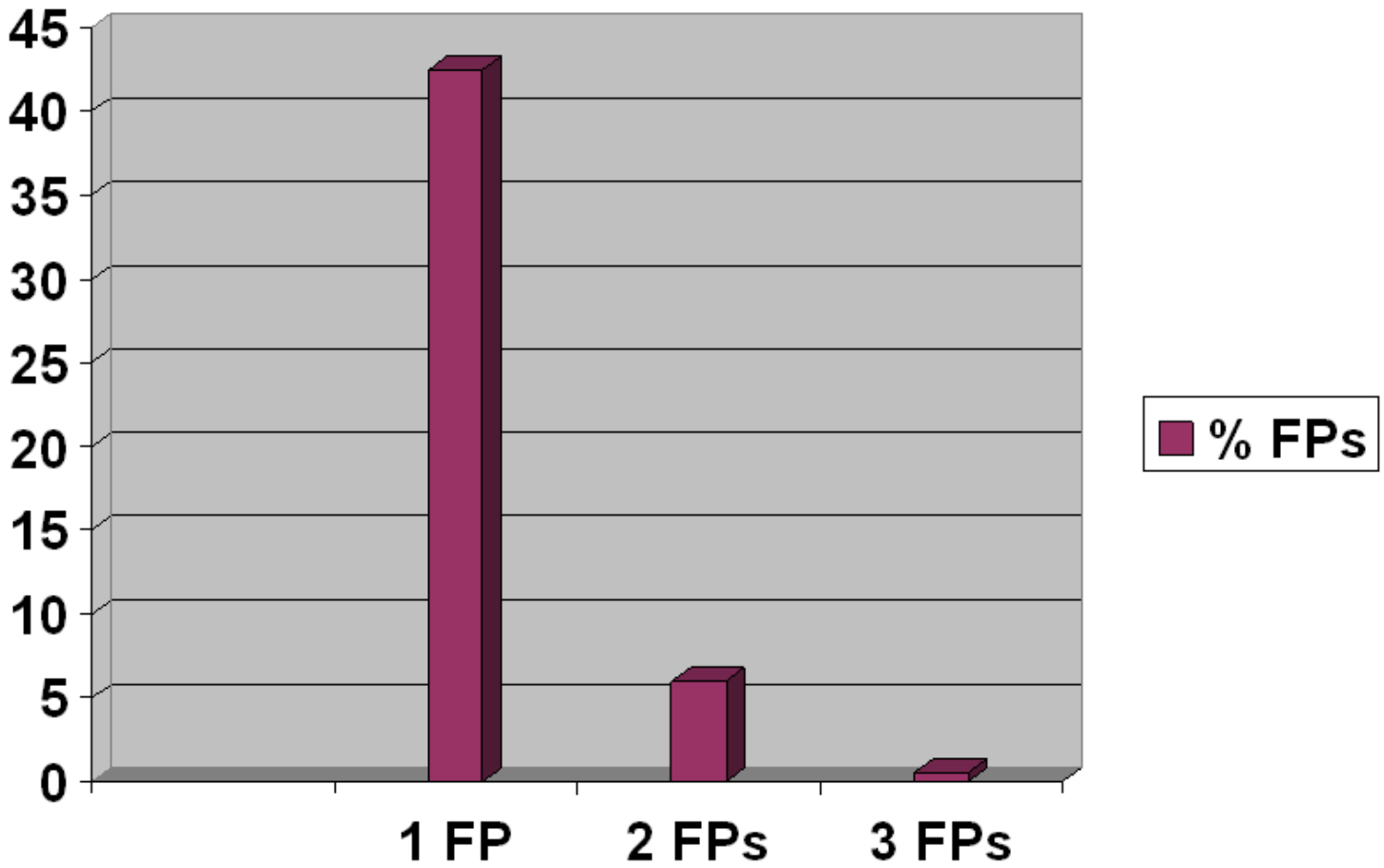
We applied the computational technique described earlier to examine:

1. if the the adults and juveniles show a differential use of pausing (filled pauses, FPs), and
2. how such a difference can be analysed and explained

Thus:

- We will now look at the 308 POS-trigrams typical for the adult (L1) speakers' syntax
  - first we look at the top 200
  - compare them to the juveniles' POS-trigrams
  - and then we look at all of them

Fig. 1: Percentage of FPs, adults top-200



Pausing

- These are the top 200 POS-trigram types which most characteristically distinguish the adults from the juveniles
  - they are significant at a $p \leq 0.05$ level
- 42.5%, 85 out of 200
  - include at least one filled pause, as in (1) and (2)
- 6%, 12 out of 200
  - include at least two filled pauses, as in (3) and (4)
- In addition, there is one trigram with only filled pauses

Pausing

(1)

	Interj	Conj(subord)	Art (def)	
politically |	uh	when	the	| liberals

(2)

	V(cop,pres,encl)	Interj	Adv(inten)	
I' |	m	ah	very	| sick

(3)

	Interj	Interj	Conj(subord)	
and |	uh	uh	because	| in the morning

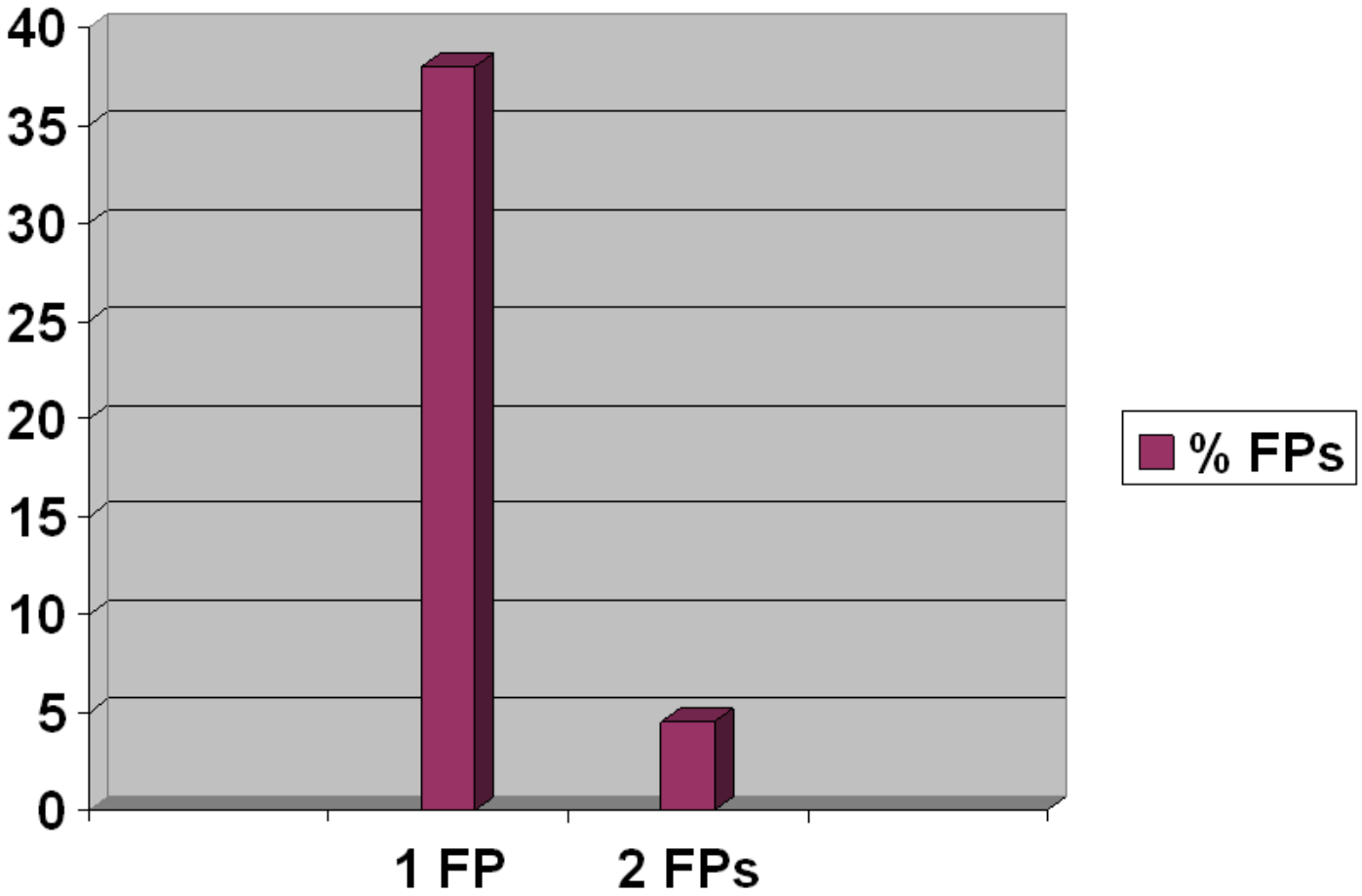
(4)

	Interj	Pron(pers, sing)	Interj	
and |	uh	I	uh	| snow-skied

Pausing

- For the juveniles there are 792 POS-trigram types in which they use the sequence of POS tags more frequently than the adults
  - Again significant at the $p \leq 0.5$ level
- But only 0.4%, 3 out of 792
  - include one filled pause
- None include more than one FP

Fig. 2: Percentage of FPs, adults all 308



Pausing

- Of all 308 POS-trigram types typical for the adults
  - this are all POS-trigrams significant at a p 0.05 level
- 38.0%, 117 out of 308
  - include at least one filled pause
- 4.5%, 14 out of 308
  - include two filled pauses
- And again there is one trigram type with FPs only

Analysis

- Both figures show the same trend. The highly skewed distribution of the filled pauses across the two groups of Finnish Australian English speakers conclusively shows that
  - the juveniles have a much more varied syntactic repertoire (measured in terms of POS-trigrams) than the adults, and
  - the adults have much more limited and idiosyncratic (ungrammatical or substandard) syntactic patterns at their disposal than the juveniles

Analysis

- The large number of filled pauses found in the adults speech as opposed to the juveniles' is in agreement with the evidence in Paananen-Porkka (2007: 234), who argues that native speakers of Finnish show longer pauses on average in English than in Finnish
- The statistically significant differential use of filled pauses by the adults can be explained in terms of the adults lesser proficiency (particularly at the level of speech planning) and, consequently, fluency of L2.

Analysis

- The elimination of all FPs from the data has little effect on the significance value for the tag sets
- The outcome of running the scripts again without the FPs showed that there are still
  - 729 statistically significant trigram types for the juveniles
  - as opposed to 220 for the adults

Analysis

- The examination of the top 200 FP-less trigram types produced by the adults showed that about 38% of the trigram types are ungrammatical, and that some of the remaining trigram types are substandard
  - (e.g. omission of an obligatory article or preposition, omission of an obligatory copula or primary verb be or have, omission of the subject, use of a redundant article with proper nouns etc.; cf. Lauttamus et al. 2007).

Conclusion

- The uneven distribution of the filled pauses across the two groups of Finnish Australian English speakers conclusively shows that
  - the adults used much more filled pauses than the juveniles, and
  - that the adults have much more limited and idiosyncratic syntactic patterns at their disposal
- The statistically significant differential use of filled pauses by the adults can be explained in terms of the adults lesser proficiency compared to that of the juveniles

Concluding Remarks

There is room for fine-tuning the method:

- Find optimum size for data-sets
- Try and evaluate with different measures

The method as is can easily be applied to many data-sets:

- Works on untagged corpora of spoken language
- Can empirically buttress theses

Software to do it and to pre-process corpora is freely available:

- the ComLinToo http://old.logilogi.org/ComLinToo

Questions

Any questions ?

Questions

Any questions ?

References

Copyleft

- Thanks to the OpenClipart archive; Carlitos for the landscape, and unknown authors for the bomb and the frogs.