

# Language Affiliation and Rank Analysis

**James W. Minett**

Language Engineering Laboratory  
City University of Hong Kong

March 30, 2004



# Outline

- **Ranks**: concept and formalization
- Using ranks to infer **linguistic affiliation**:
  - Scenario #1 — an **idealized** case
  - Scenario #2 — a *somewhat more realistic* case
- **Summary**  
and some prospects for **more realistic** modeling of **linguistic practice**



# Ranks — Concept

- “**Universal Rank Analysis**” (Chen, 1995):  
Swadesh (1952, 1955) basic words are partitioned into **2 ranks**:
  - **Rank 1**: Swadesh **100-word list**, found to have:
    - **higher probability of retention**
    - **lower probability** of being replaced by **borrowing** than:
  - **Rank 2**: remaining words of the Swadesh **200-word list**
- “**Relativistic Rank Analysis**” (Chen & He, 2002):  
Arbitrary set of words are partitioned into **2 ranks**:
  - **Rank 1**: with **high correspondence rate** across language set
  - **Rank 2**: with **low correspondence rate** across language set

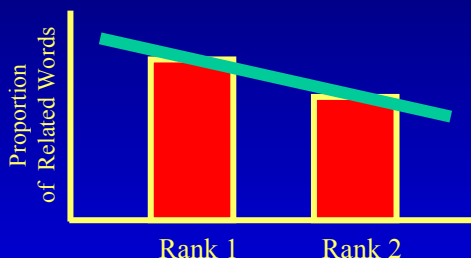
e.g.

	Northern Tai	Central Tai	SW Tai
<b>Rank 1</b>	‘dog’ <b>ma<sup>1</sup></b>	<b>ma<sup>1</sup></b>	<b>ma<sup>1</sup></b>
<b>Rank 2</b>	‘sheep’ ji:ꨀ <sup>2</sup>	be <sup>3</sup>	me <sup>3</sup>

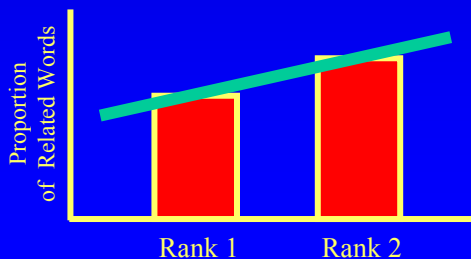


## Ranks — Concept

- Languages having a **genetic affiliation** are expected to have a **greater proportion** of related words in **Rank 1** than in Rank 2



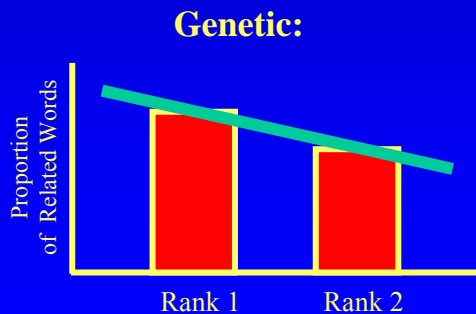
- Languages having **no genetic affiliation (contact)** are expected to have a **greater proportion** of related words in **Rank 2** than in Rank 1.





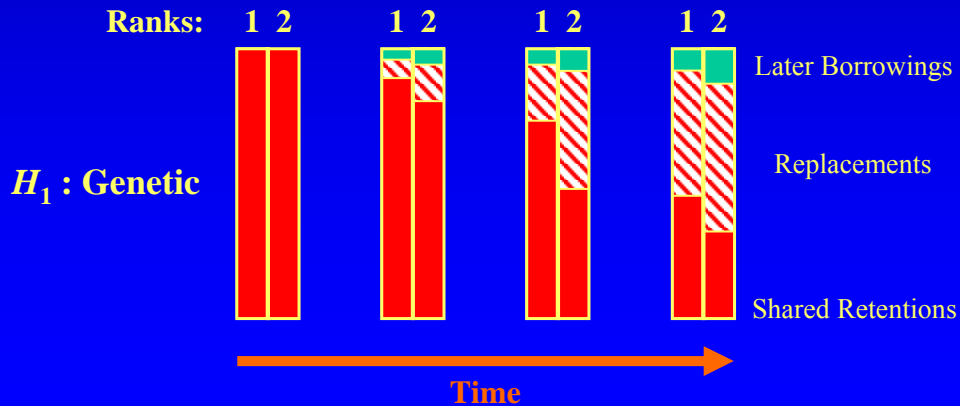
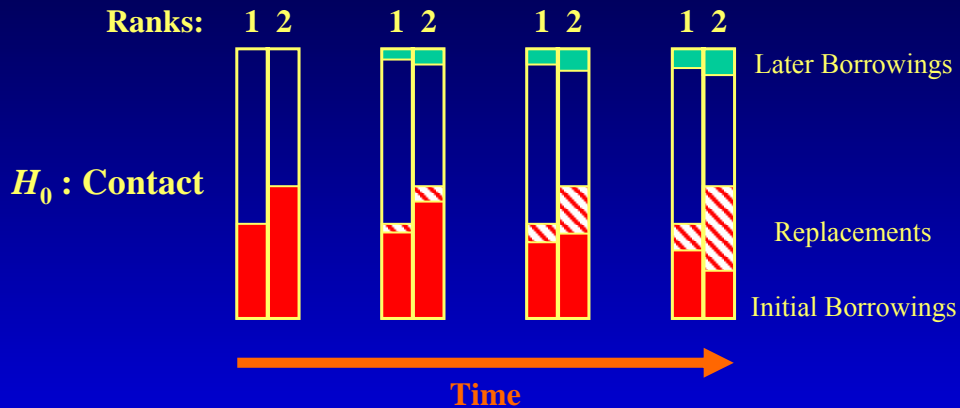
## Ranks — Decision Rule

- Proportion of Words in **Rank 1**  $>$  Proportion of Words in **Rank 2**  
 $\Rightarrow$  **Genetic** Affiliation
- Proportion of Words in **Rank 1**  $\leq$  Proportion of Words in **Rank 2**  
 $\Rightarrow$  Affiliation due to **Contact**





# Ranks — Evolution





## Ranks — Formalization

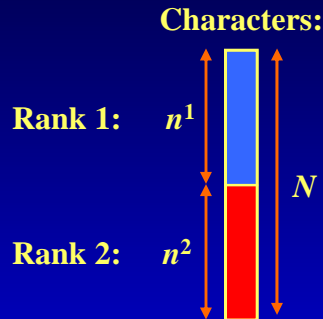
- Consider a set of (lexical) **characters**

$$\mathbf{C} = \{\mathbf{c}^q : q = 1, \dots, N\}$$

that have been partitioned into 2 ranks:

$$\mathbf{R}^1 = \{\mathbf{c}^q : q = 1, \dots, n^1\}$$

$$\mathbf{R}^2 = \{\mathbf{c}^q : q = n^1+1, \dots, n^2\}$$



- Consider 2 languages,  $L_1$  &  $L_2$ , whose affiliation is to be tested. Count number of **related words** in  $L_1$  &  $L_2$  in each rank:  $n_{12}^1$  &  $n_{12}^2$
- Define the **rank difference**:  $\delta_V = n_{12}^1/n^1 - n_{12}^2/n^2$

$$\text{e.g. } \delta_V = 35/100 - 18/100 = +17\%$$

The languages are declared to be **genetically related** when  $\delta_V > 0$



- The appropriate question to focus on is not

“**does** Rank Analysis work?”

but

“**how often** / **when** does it work?”

and, particularly,

“**how often** / **when** does it work significantly **better than chance**?”

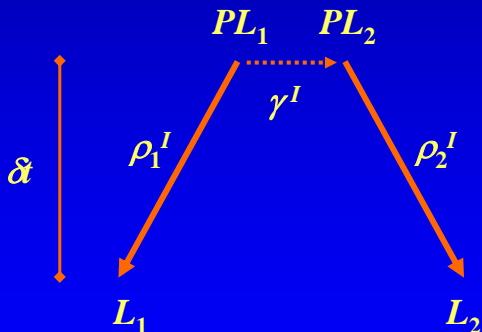
- Now consider performance for 2 scenarios: ...



# Scenario #1

- Idealized case:— no intermediate borrowing; no errors; homogeneous:

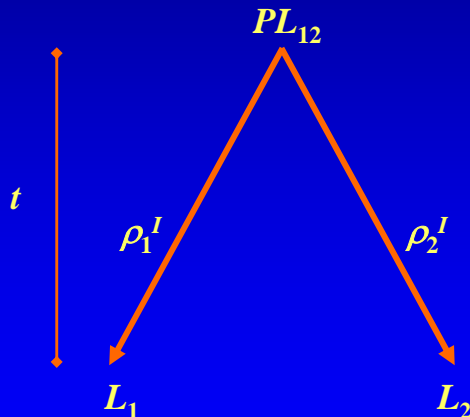
$H_0$  : Contact



$$p_0^I = \Pr(c_1^I = c_2^I \mid H_0) = \gamma (\rho_1^I)^{\delta t} (\rho_2^I)^{\delta t}$$

i.e. Pr (related words | Contact)

$H_1$  : Genetic



$$p_1^I = \Pr(c_1^I = c_2^I \mid H_1) = (\rho_1^I)^t (\rho_2^I)^t$$

i.e. Pr (related words | Genetic)



## Scenario #1 — Calculation

- The languages are declared to be **genetically related** when  $\delta\nu > 0$
- Under  $H_0$ : Contact:

$$\begin{aligned}\Pr(\delta\nu > 0 \mid H_0) &= \Pr\left(\frac{1}{n^1} \text{Bin}(n^1, p_0^1) > \frac{1}{n^2} \text{Bin}(n^2, p_0^2)\right) \\ &\approx \Pr\left(\frac{1}{n^1} \text{N}(n^1 p_0^1, n^1 p_0^1 (1 - p_0^1)) > \frac{1}{n^2} \text{N}(n^2 p_0^2, n^2 p_0^2 (1 - p_0^2))\right) \\ &= \Pr\left(\text{N}(p_0^2 - p_0^1, p_0^2(1 - p_0^2)/n^2 + p_0^1(1 - p_0^1)/n^1) < 0\right) \\ &= \Phi\left(\frac{p_0^2 - p_0^1}{\sqrt{p_0^2(1 - p_0^2)/n^2 + p_0^1(1 - p_0^1)/n^1}}\right)\end{aligned}$$

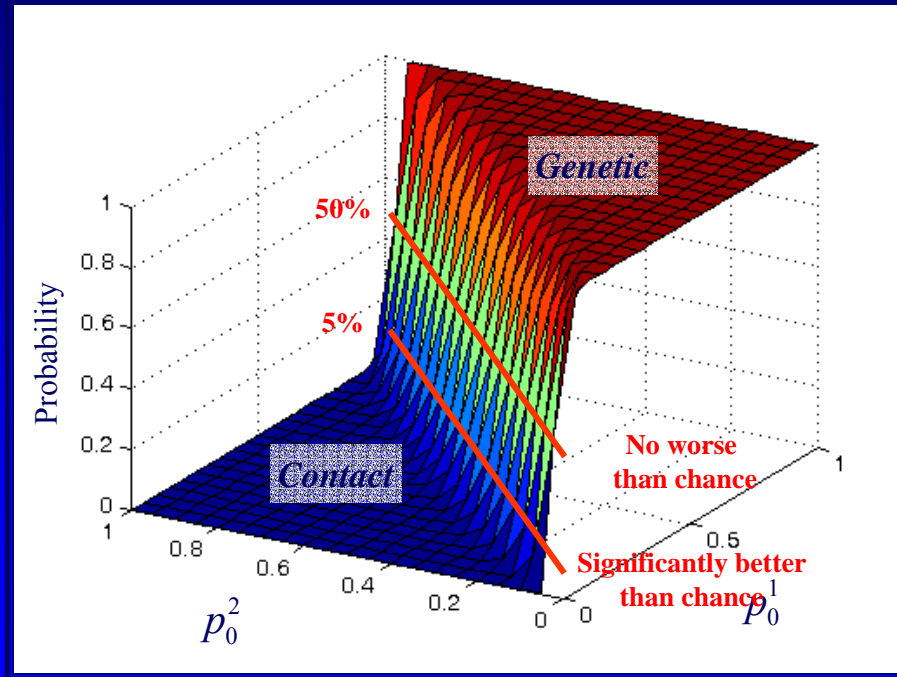
which allows the performance to be **analyzed numerically**

(Similar expression can be obtained for hypothesis  $H_1$ : Genetic)



## Scenario #1 — False Alarm Rate

- Probability distribution:  $\Pr(\delta v > 0 | H_0)$  ( $n^1 = n^2 = 100$ ):



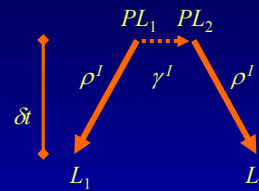
$$p^I = \Pr(\text{Related words in Rank } I)$$



## Scenario #1 — Examples

- For  $\delta=2$ ,  $\gamma^1=5\%$ ,  $\gamma^2=15\%$ ,  $\rho^1=90\%$ ,  $\rho^2=80\%$ :

$$p_0^1=3.3\%, p_0^2=6.1\%$$



hence **<<50% probability** of incorrectly declaring **genetic** affiliation

- But for  $\delta=2$ ,  $\gamma^1=5\%$ ,  $\gamma^2=10\%$ ,  $\rho^1=95\%$ ,  $\rho^2=80\%$ : :

$$p_0^1=4.1\%, p_0^2=4.1\%$$

hence **~50% probability** of incorrectly declaring **genetic** affiliation  
— **no better than chance!**



## Scenario #1 — Limits of Method

- Analysis can be made easier by **non-dimensionalizing** the system:

$$\text{Writing } \rho_0^2 = r\rho_0^1 \text{ and } g\gamma_0^2 = \gamma_0^1, \quad \frac{p_0^2}{p_0^1} = \frac{\gamma_0^2 (\rho_0^2)^{2\delta}}{\gamma_0^1 (\rho_0^1)^{2\delta}} = \frac{r^{2\delta}}{g}$$

- False Alarm Rate < 50% when  $p_0^2 \leq p_0^1$ , i.e.  $g \leq r^{2\delta}$
- For example, for  $r \geq 80\%$  and  $\delta \leq 4$

$$p_0^2 \geq p_0^1 \text{ when } g \leq 16.8\%$$

e.g.  $\delta \leq 4$ ,  $\rho^1 = 90\%$ ,  $\rho^2 \geq 72\%$ ,  $\gamma^1 \leq 3.3\%$ ,  $\gamma^2 = 20\%$

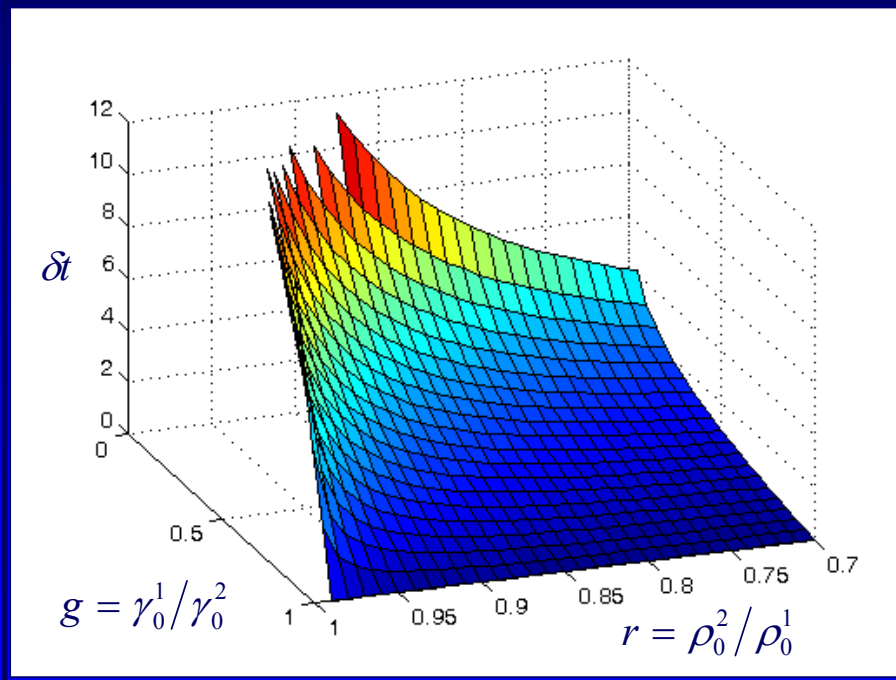
e.g.  $\delta \leq 2$ ,  $\rho^1 = 90\%$ ,  $\rho^2 \geq 72\%$ ,  $\gamma^1 \leq 8.2\%$ ,  $\gamma^2 = 20\%$

(Bounds on performance for **time-varying** retention rates can be processed in this way)



## Scenario #1 — Limits of Method

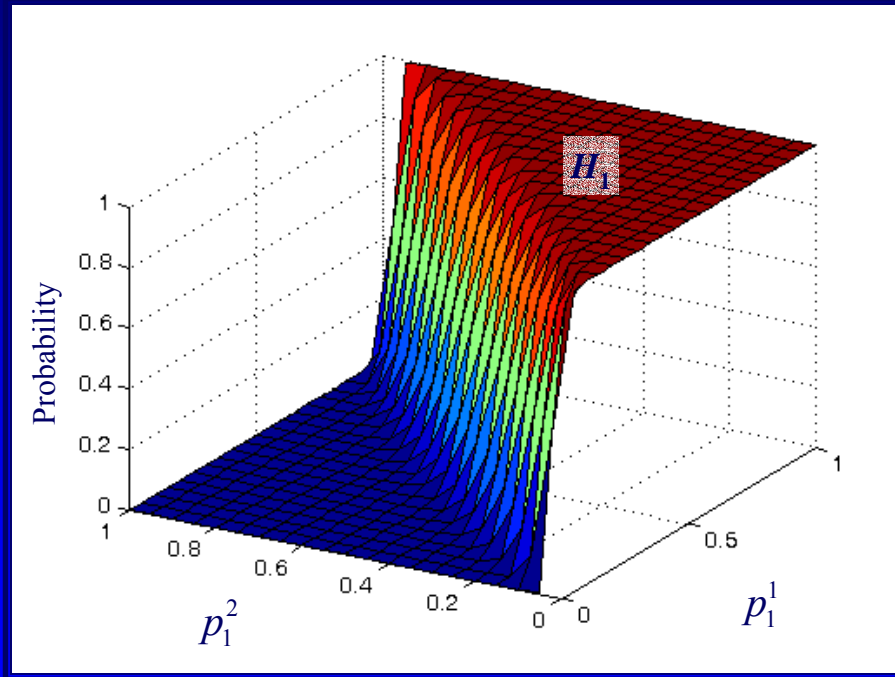
- Rank method attains **< 50% error** for  $(r, g, \delta t)$  below the threshold:





## Scenario #1 — Detection Rate

- Detection rate:  $\Pr(\delta \mathbf{v} > 0 | H_1)$  ( $n^1 = n^2 = 100$ ):  $p_1^t = (\rho_1^t)' (\rho_2^t)'$



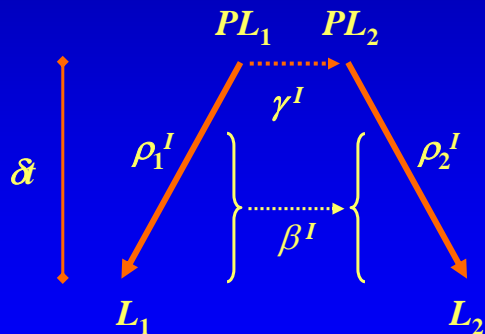
Detection rate better than chance virtually **guaranteed**



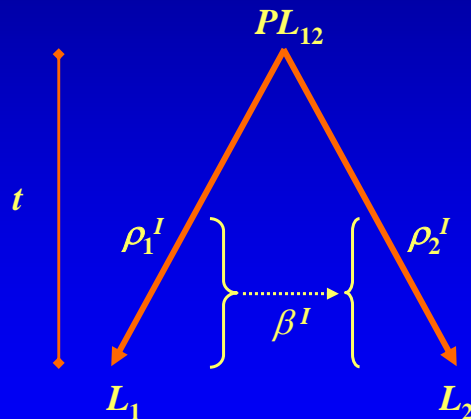
## Scenario #2

- *More* realistic case:— **intermediate borrowing**; no errors; homogeneous:

$H_0$  : Contact

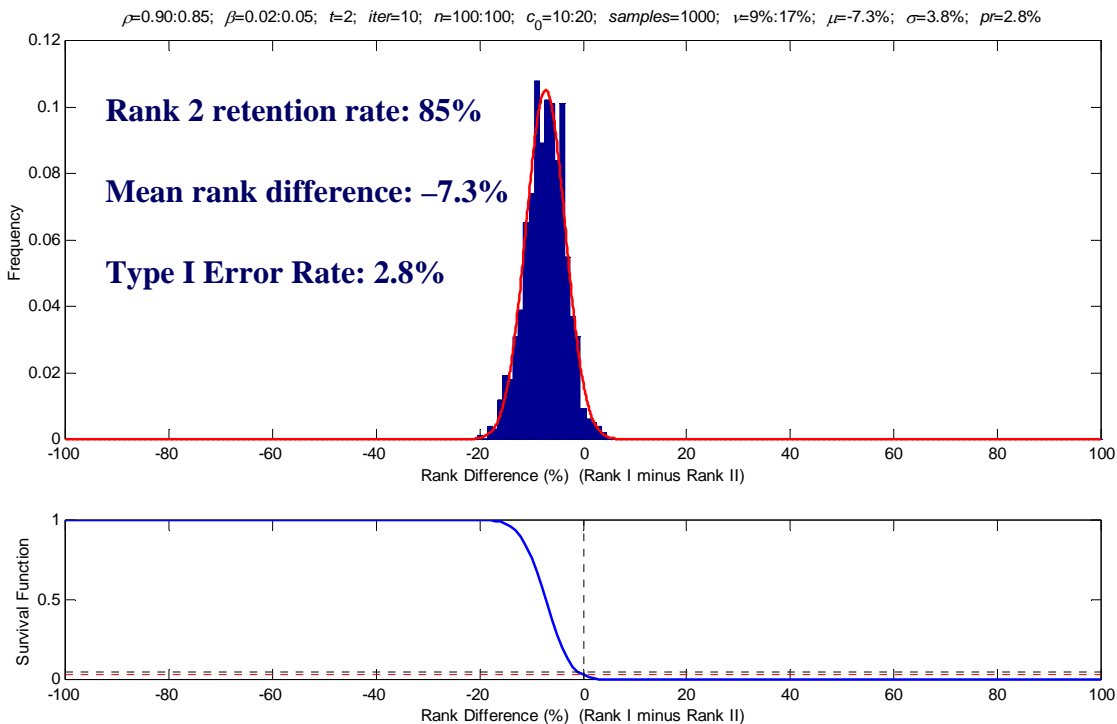


$H_1$  : Genetic



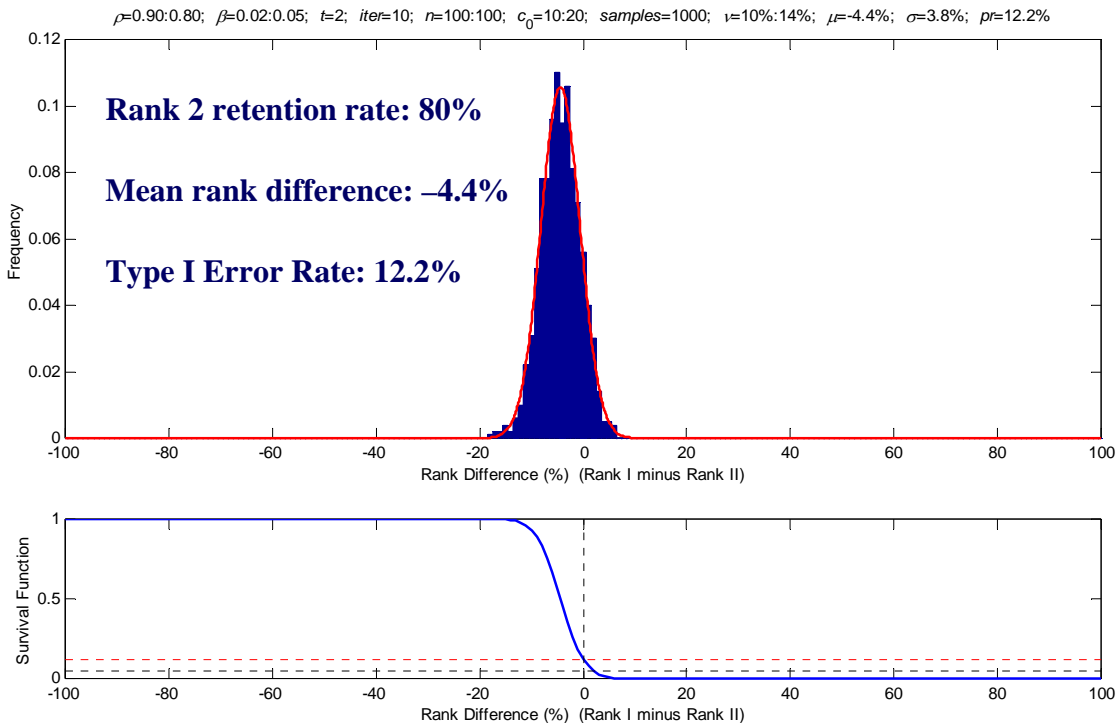


## Scenario #2 — Simulation 1 ( $H_0$ : Contact)



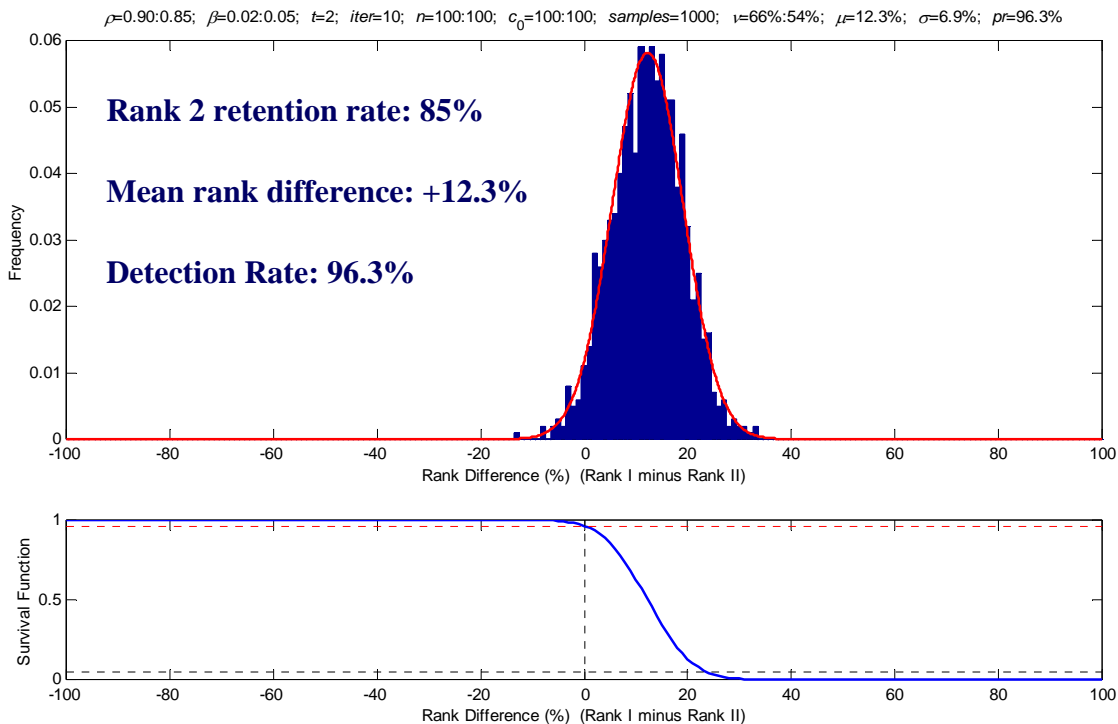


## Scenario #2 — Simulation 2 ( $H_0$ : Contact)



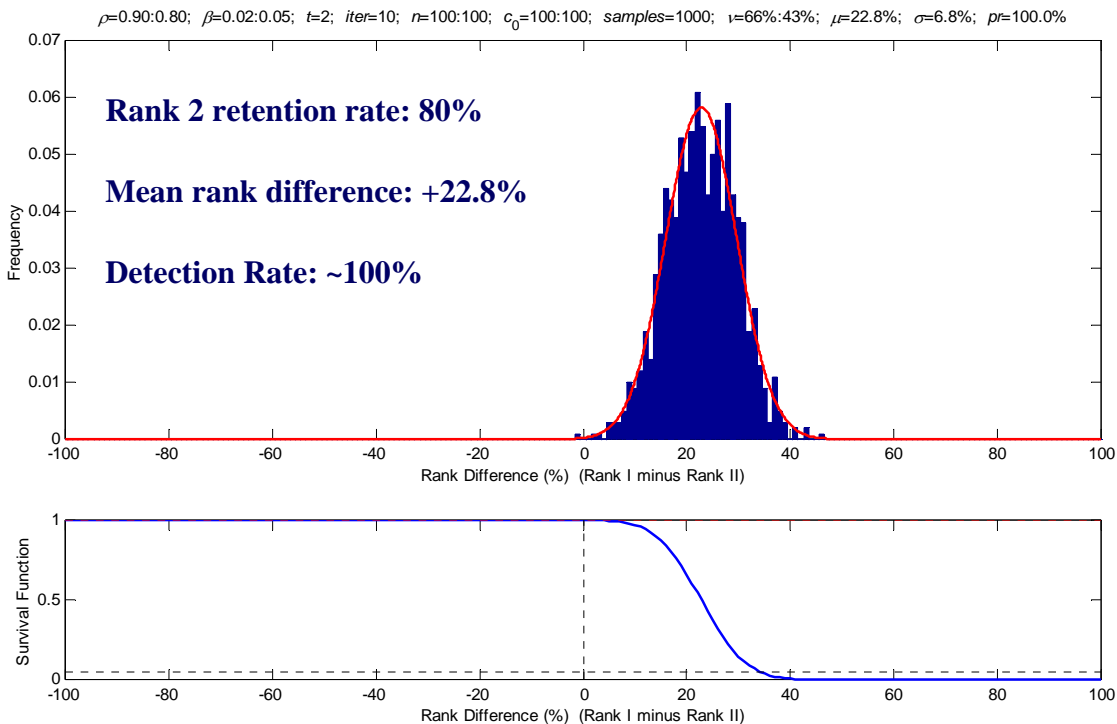


# Scenario #2 — Simulation 1 ( $H_1$ : Genetic)



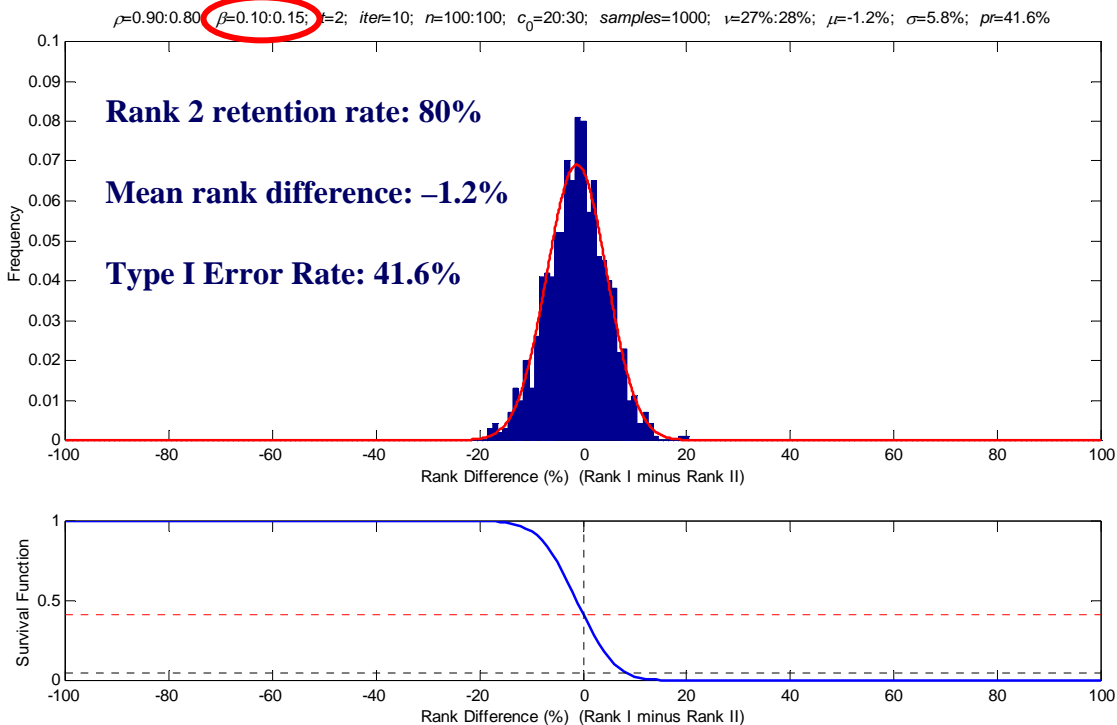


## Scenario #2 — Simulation 2 ( $H_1$ : Genetic)



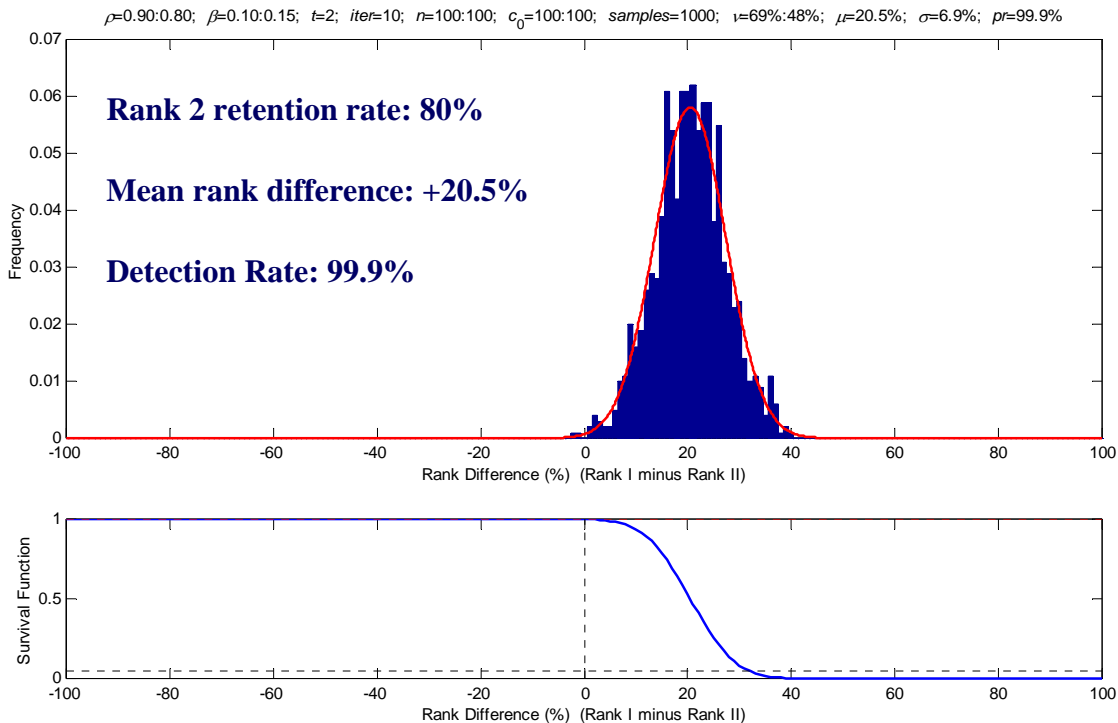


# Scenario #3 — It Can Sometimes Go Wrong!





# Scenario #3 — But At Least Detection Rate Remains High





## What About Heterogeneous Characters?

- Different characters have **different retention rates** and **different borrowing rates**
- Determine loose **upper** and **lower bounds** on **performance** by specifying **upper** and **lower bounds** on retention and borrowing rates

- For example, what if:

$$\begin{array}{ll} \frac{1}{2} \leq \alpha \leq 2 & 2 \leq t \leq 5, \\ 75\% \leq \rho^1 \leq 90\% & 65\% \leq \rho^2 \leq 80\%, \\ 0\% \leq \gamma^1 \leq 10\% & 5\% \leq \gamma^2 \leq 20\%, \\ 0\% \leq \beta^1 \leq 10\% & 5\% \leq \beta^2 \leq 20\% \end{array}$$

Oh dear!

$$0\% \leq P_{fa} \leq 100\% \quad \text{and} \quad 0\% \leq P_d \leq 100\%$$



## Summary

- The rank method can be analyzed using **statistical methods**
- In ideal situations (no intermediate contact), **detection rate  $\gg 50\%$** ; but **low false alarm rates** can only be attained for **small time depth**  
— **precise bounds to be determined**
- Analysis for somewhat more realistic situations (intermediate contact) undertaken by **simulation**
- **High detection rates** maintained; **false alarms** remain **problematic**
- Still to analyze:  
**heterogeneous characters** and **stratification ...**



## Prospects for More Realistic Modeling of Linguistic Practice?

- **Stratification** and Affiliation:  
By **stratifying** a set of lexical characters into a set of **ordered strata**, only the **earliest layer** of which can possibly represent **genetic** retentions, many **borrowings** can potentially be **excised**
- So the **rank method** should work **better**, shouldn't it?
- Well, probably ...  
provided that:
  - sufficiently **many related words remain** for significance;
  - the linguist has **stratified** reasonably **accurately**;
  - the linguist *has* recovered the **oldest layer**
- **Performance remains to be tested** — work with Chen??



## References

- Chen, Baoya & He, Fang. 2002. “Relativistic rank analysis of kernel consistent corresponding words between Chinese and Kam-Tai”. *Linguistics of the Tibeto-Burman Area* 25.2:195–224.
- Chen, Baoya. 1995. “On the original relationship between Chinese and Kam-Tai”. *Linguistics of the Tibeto-Burman Area* 18.1:149–171.