# Informatics 1B
# Data and Analysis: Tutorial 4
# Chi-Square tests and Vector Space Models

Manuel Marques Pita and Frank Keller

February 21, 2007

## 1 Comparing word usage in film translations

The film **A Very Long Engagement** directed by Jean Jeunet has been already translated into dozens of languages. Translating all the dialogues into Portuguese was an annoying and difficult task as it had to be done three times: one for the Portuguese audience, another for the Brazilians and a last one for the Angolans. This happened because, even though these three countries speak Portuguese, the use of certain words varies a lot. There are many interesting cases of words which have meaning variations. One of them is the word *rapariga* which in Portuguese means basically girl. However, in Brazilian Portuguese this word is used to refer to a prostitute. In Angola, this word can have either of these two meanings depending on the context.

Translating films into different languages costs a lot of money and takes a long time. If the goal was to save money and time, they could have used the word menina which means girl in these three languages. One of the arguments against this idea was that in Portugal and Angola it sounds funny to refer to a girl as *menina* as it is used sometimes in sarcasm. One question to be considered in this context is the following: After all the work, is there a significant variance in the use of the words *menina* and *rapariga* in these three translations of this film?

*Chi squared* tests are covered in Lecture 9 of Data and Analysis. The section of the notes you need to read for this tutorial starts on page 58.

**Question 1**

If the goal is to explore the possibility of using only one (or two) translators next time and we want to test the variance using a chi-square test, what should the *Null Hypothesis* be?

**Question 2**

Figure 1 shows the number of times the words *rapariga* and *menina* were used in the translation dialogues (which were done by native speakers from each of the three countries who are also fluent in French). Given this table, calculate the *expected values*, $\mathrm{E}_{i,j}$, for this example.

| | *rapariga* | *menina* | Totals |
|---|---|---|---|
| **Portuguese** | 47 | 32 | 79 |
| **Brazilian** | 21 | 28 | 49 |
| **Angolan** | 45 | 39 | 84 |
| Totals | 113 | 99 | 212 |

**Figure 1**: Frequency of use of words *rapariga* and *menina* in translated dialogues

**Question 3**

Calculate the *chi-squared*, $\chi^2$, value for this example.

**Question 4**

What needs to be done now that the chi-squared value has been calculated in order to accept or reject the Null Hypothesis (Hint: the corresponding critical value for a significance level of $p<0.05$ with two degrees of freedom is 5.991).

## 2 A basic IR example for the Portuguese translations

In the previous question you analysed how significantly different the usage of the words *rapariga* and *menina* are in three versions of the Portuguese language for the translation of a film originally in French.

| | *rapariga* | *menina* |
|---|---|---|
| **Portuguese** | 5 | 3 |
| **Brazilian** | 2 | 3 |
| **Angolan** | 4 | 8 |

**Figure 2**: Frequency of use of *rapariga* and *menina* during first 25 minutes of translated dialogues.

Our next task is to help the consultant deciding on budgets for translations into Portuguese of other French films. For this, you need to report on the *usage in context* of those words that have been used as the argument for having several Portuguese translations. In this example, we will only concentrate again on the the words rapariga and menina. We want to formulate queries such as Q1=*rapariga* to see how this word is used in different contexts. We could probably do this over the three available translations, but our consultant is in a big hurry and therefore she needs an answer as soon as possible. If we are to analyse only one document in which we can maximise the chances of looking into the different contexts in which this word is used, which document should we go for?

In order to find out, we will work with a *Vector Space Model* for this problem. You will need to look into your lectures notes starting on page 65 for this. This material may not have been covered by the time you do this tutorial. To simplify your work, we will work with the first 25 minutes of the dialogues translated in the three languages. The corresponding frequencies for the words we are studying is shown in Figure 2.

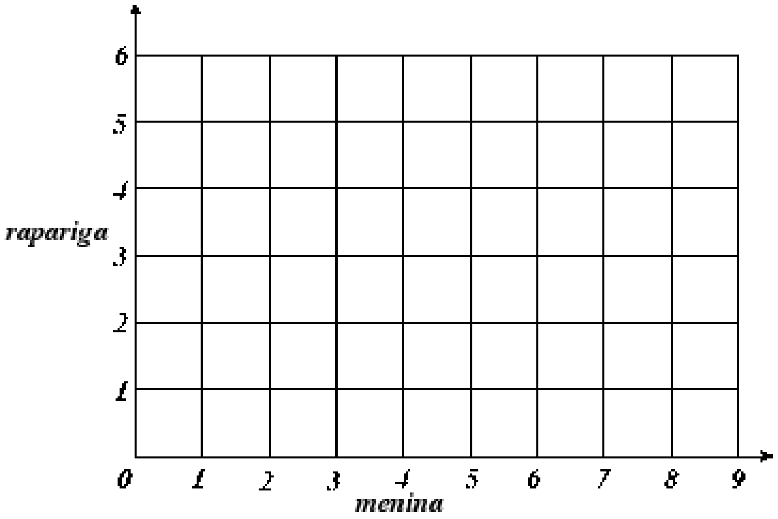

**Figure 3**: Graph for representing vectors in the space model for the Portuguese problem

**Question 5**

Using the graph provided in Figure 3, draw the document vectors.

**Question 6** For the query (Q1 = rapariga AND menina), draw the query vector in your vector space. Intuitively, which document would you choose as the one to rank at the top?

**Question 7**

Calculate the *cosines* between each document and the query Q1 using the formula II.5 in page 68 of the notes. How are the documents ranked according to this similarity measure? Does your initial intuition agree with the results?

**Question 8**

Rank the available documents (using again VSM and the cosine similarity measure) for the query Q2 = rapariga.