

# Linguistic contributions towards the selection of relevant information in clinical interviews

Marcos Lopes<sup>1</sup>, Viviana Giampaoli<sup>2</sup>, Maria José Baraldi<sup>1</sup>,  
Alexandre Suzuki<sup>3</sup>, Elisete Aubin<sup>2</sup>, Alfredo José Mansur<sup>3</sup>

<sup>1</sup>Linguistics Department – FFLCH – USP

<sup>2</sup>Statistics Department – IME – USP

<sup>3</sup>Heart Institute – FM – USP

Humboldt Kolleg  
Limits and Interfaces in Science  
11/29/2009



# The problem

- How worthy are the contents in patients' speech comparing to clinical and laboratorial methods of diagnostics?
- Many studies have addressed this question applying written forms and scales that should be filled by patients according to their perception of symptoms (Cole *et al.* 1999; Appels *et al.* 1987; van Diest 1990)
- Results would then be correlated to clinical and laboratorial data in order to evaluate any correspondances between them

# The problem

- How worthy are the contents in patients' speech comparing to clinical and laboratorial methods of diagnostics?
- Many studies have addressed this question applying written forms and scales that should be filled by patients according to their perception of symptoms (Cole *et al.* 1999; Appels *et al.* 1987; van Diest 1990)
- Results would then be correlated to clinical and laboratorial data in order to evaluate any correspondances between them

# The problem

- How worthy are the contents in patients' speech comparing to clinical and laboratorial methods of diagnostics?
- Many studies have addressed this question applying written forms and scales that should be filled by patients according to their perception of symptoms (Cole *et al.* 1999; Appels *et al.* 1987; van Diest 1990)
- Results would then be correlated to clinical and laboratorial data in order to evaluate any correspondances between them

# The problems with forms

- Such methodology, even if it proves worthwhile in their original countries, would hardly be applied in Brazil because of:
  - The relatively low degree of literacy of a great part of the population
  - The fact that poor Brazilians are far more used to answering to oral questions, not to written ones — specially not abstract scales
  - The social heterogeneity of patients, which makes them produce very disparate interpretations of clinical questions

# The problems with forms

- Such methodology, even if it proves worthwhile in their original countries, would hardly be applied in Brazil because of:
  - The relatively low degree of literacy of a great part of the population
  - The fact that poor Brazilians are far more used to answering to oral questions, not to written ones — specially not abstract scales
  - The social heterogeneity of patients, which makes them produce very disparate interpretations of clinical questions

# The problems with forms

- Such methodology, even if it proves worthwhile in their original countries, would hardly be applied in Brazil because of:
  - The relatively low degree of literacy of a great part of the population
  - The fact that poor Brazilians are far more used to answering to oral questions, not to written ones — specially not abstract scales
  - The social heterogeneity of patients, which makes them produce very disparate interpretations of clinical questions

# The problems with forms

- Such methodology, even if it proves worthwhile in their original countries, would hardly be applied in Brazil because of:
  - The relatively low degree of literacy of a great part of the population
  - The fact that poor Brazilians are far more used to answering to oral questions, not to written ones — specially not abstract scales
  - The social heterogeneity of patients, which makes them produce very disparate interpretations of clinical questions

# Objectives

- To establish quantitative criteria of analysis from natural language speech in order to allow correlations between linguistic and clinical data
- First hypothesis: Linguistic contents are not trivial — it adds new information to clinical data
- Second hypothesis: Linguistic contents are not informationnally homogeneous; relevance varies

# Objectives

- To establish quantitative criteria of analysis from natural language speech in order to allow correlations between linguistic and clinical data
- First hypothesis: Linguistic contents are not trivial — it adds new information to clinical data
- Second hypothesis: Linguistic contents are not informationnally homogeneous; relevance varies

# Objectives

- To establish quantitative criteria of analysis from natural language speech in order to allow correlations between linguistic and clinical data
- First hypothesis: Linguistic contents are not trivial — it adds new information to clinical data
- Second hypothesis: Linguistic contents are not informationnally homogeneous; relevance varies

# Subjects and data

- 266 heart failure patients from the Heart Institute of the University of São Paulo
- Patients were asked seven questions concerning their general health and perception of symptoms
- Interviews were recorded and transcribed to text files that were registered in a relational database
- We have selected one out of those seven clinical questions for the purpose of this investigation:

Do you have any difficulties to breathe?

- Text data were analysed by two linguists who classified the contents according to lexical criteria

# Subjects and data

- 266 heart failure patients from the Heart Institute of the University of São Paulo
- Patients were asked seven questions concerning their general health and perception of symptoms
- Interviews were recorded and transcribed to text files that were registered in a relational database
- We have selected one out of those seven clinical questions for the purpose of this investigation:

Do you have any difficulties to breathe?

- Text data were analysed by two linguists who classified the contents according to lexical criteria

# Subjects and data

- 266 heart failure patients from the Heart Institute of the University of São Paulo
- Patients were asked seven questions concerning their general health and perception of symptoms
- Interviews were recorded and transcribed to text files that were registered in a relational database
- We have selected one out of those seven clinical questions for the purpose of this investigation:

Do you have any difficulties to breathe?

- Text data were analysed by two linguists who classified the contents according to lexical criteria

# Subjects and data

- 266 heart failure patients from the Heart Institute of the University of São Paulo
- Patients were asked seven questions concerning their general health and perception of symptoms
- Interviews were recorded and transcribed to text files that were registered in a relational database
- We have selected one out of those seven clinical questions for the purpose of this investigation:

Do you have any difficulties to breathe?

- Text data were analysed by two linguists who classified the contents according to lexical criteria

# Subjects and data

- 266 heart failure patients from the Heart Institute of the University of São Paulo
- Patients were asked seven questions concerning their general health and perception of symptoms
- Interviews were recorded and transcribed to text files that were registered in a relational database
- We have selected one out of those seven clinical questions for the purpose of this investigation:

Do you have any difficulties to breathe?

- Text data were analysed by two linguists who classified the contents according to lexical criteria

# Subjects and data

- 266 heart failure patients from the Heart Institute of the University of São Paulo
- Patients were asked seven questions concerning their general health and perception of symptoms
- Interviews were recorded and transcribed to text files that were registered in a relational database
- We have selected one out of those seven clinical questions for the purpose of this investigation:

Do you have any difficulties to breathe?

- Text data were analysed by two linguists who classified the contents according to lexical criteria

# Clinical variables

- age
- body mass index
- blood pressure
- Laboratorial data:
  - hemoglobin
  - hematocrit
  - serum cholesterol
  - triglycerides
  - creatinine
  - serum glucose

# Clinical variables

- age
- body mass index
- blood pressure
- Laboratorial data:
  - hemoglobin
  - hematocrit
  - serum cholesterol
  - triglycerides
  - creatinine
  - serum glucose

# Clinical variables

- age
- body mass index
- blood pressure
- Laboratorial data:
  - hemoglobin
  - hematocrit
  - serum cholesterol
  - triglycerides
  - creatinine
  - serum glucose

# Clinical variables

- age
- body mass index
- blood pressure
- Laboratorial data:
  - hemoglobin
  - hematocrit
  - serum cholesterol
  - triglycerides
  - creatinine
  - serum glucose

# Clinical variables

- age
- body mass index
- blood pressure
- Laboratorial data:
  - hemoglobin
  - hematocrit
  - serum cholesterol
  - triglycerides
  - creatinine
  - serum glucose

# Clinical variables

- age
- body mass index
- blood pressure
- Laboratorial data:
  - hemoglobin
  - hematocrit
  - serum cholesterol
  - triglycerides
  - creatinine
  - serum glucose

# Clinical variables

- age
- body mass index
- blood pressure
- Laboratorial data:
  - hemoglobin
  - hematocrit
  - serum cholesterol
  - triglycerides
  - creatinine
  - serum glucose

# Clinical variables

- age
- body mass index
- blood pressure
- Laboratorial data:
  - hemoglobin
  - hematocrit
  - serum cholesterol
  - triglycerides
  - creatinine
  - serum glucose

# Clinical variables

- age
- body mass index
- blood pressure
- Laboratorial data:
  - hemoglobin
  - hematocrit
  - serum cholesterol
  - triglycerides
  - creatinine
  - serum glucose

# Clinical variables

- age
- body mass index
- blood pressure
- Laboratorial data:
  - hemoglobin
  - hematocrit
  - serum cholesterol
  - triglycerides
  - creatinine
  - serum glucose

# Linguistic variables

## Medium Range Thematization (MRT) of descriptions of breathlessness

- MRT is a semantic evaluation of lexical terms
- It represents a classification of words according to semantical classes:
  - "part of the human body"
  - "physical activities"
  - "sleep position"
  - "psychological factors"
  - "medicines"
  - "environment factors"
- ... and their combinations, adding up to 20 categories of responses or **linguistic clusterings**

# Linguistic variables

## Medium Range Thematization (MRT) of descriptions of breathlessness

- MRT is a semantic evaluation of lexical terms
- It represents a classification of words according to semantical classes:
  - "part of the human body"
  - "physical activities"
  - "sleep position"
  - "psychological factors"
  - "medicines"
  - "environment factors"
- ... and their combinations, adding up to 20 categories of responses or **linguistic clusterings**

# Linguistic variables

## Medium Range Thematization (MRT) of descriptions of breathlessness

- MRT is a semantic evaluation of lexical terms
- It represents a classification of words according to semantical classes:
  - “part of the human body”
  - “physical activities”
  - “sleep position”
  - “psychological factors”
  - “medicines”
  - “environment factors”
- ... and their combinations, adding up to 20 categories of responses or **linguistic clusterings**

# Linguistic variables

## Medium Range Thematization (MRT) of descriptions of breathlessness

- MRT is a semantic evaluation of lexical terms
- It represents a classification of words according to semantical classes:
  - “part of the human body”
  - “physical activities”
  - “sleep position”
  - “psychological factors”
  - “medicines”
  - “environment factors”
- ... and their combinations, adding up to 20 categories of responses or **linguistic clusterings**

# Linguistic variables

## Medium Range Thematization (MRT) of descriptions of breathlessness

- MRT is a semantic evaluation of lexical terms
- It represents a classification of words according to semantical classes:
  - “part of the human body”
  - “physical activities”
  - “sleep position”
  - “psychological factors”
  - “medicines”
  - “environment factors”
- ... and their combinations, adding up to 20 categories of responses or **linguistic clusterings**

# Linguistic variables

## Medium Range Thematization (MRT) of descriptions of breathlessness

- MRT is a semantic evaluation of lexical terms
- It represents a classification of words according to semantical classes:
  - “part of the human body”
  - “physical activities”
  - “sleep position”
  - “psychological factors”
  - “medicines”
  - “environment factors”
- ... and their combinations, adding up to 20 categories of responses or **linguistic clusterings**

# Linguistic variables

## Medium Range Thematization (MRT) of descriptions of breathlessness

- MRT is a semantic evaluation of lexical terms
- It represents a classification of words according to semantical classes:
  - “part of the human body”
  - “physical activities”
  - “sleep position”
  - “psychological factors”
  - “medicines”
  - “environment factors”
- ... and their combinations, adding up to 20 categories of responses or **linguistic clusterings**

# Linguistic variables

## Medium Range Thematization (MRT) of descriptions of breathlessness

- MRT is a semantic evaluation of lexical terms
- It represents a classification of words according to semantical classes:
  - “part of the human body”
  - “physical activities”
  - “sleep position”
  - “psychological factors”
  - “medicines”
  - “environment factors”
- ... and their combinations, adding up to 20 categories of responses or **linguistic clusterings**

# Linguistic variables

## Medium Range Thematization (MRT) of descriptions of breathlessness

- MRT is a semantic evaluation of lexical terms
- It represents a classification of words according to semantical classes:
  - “part of the human body”
  - “physical activities”
  - “sleep position”
  - “psychological factors”
  - “medicines”
  - “environment factors”
- ... and their combinations, adding up to 20 categories of responses or **linguistic clusterings**

# Linguistic variables

## Medium Range Thematization (MRT) of descriptions of breathlessness

- MRT is a semantic evaluation of lexical terms
- It represents a classification of words according to semantical classes:
  - “part of the human body”
  - “physical activities”
  - “sleep position”
  - “psychological factors”
  - “medicines”
  - “environment factors”
- ... and their combinations, adding up to 20 categories of responses or **linguistic clusterings**

# Resetting the problem

- Both linguistic and clinical data for the sample population can be organized in groups
- Let  $L$  and  $C$  be the names of those groups
- Questions in our Objectives could be rephrased as:

How much information is added by  $L$  to  $C$ ?

Knowing the organization of  $L$ , what could one say about  $C$ ?

# Resetting the problem

- Both linguistic and clinical data for the sample population can be organized in groups
- Let  $L$  and  $C$  be the names of those groups
- Questions in our Objectives could be rephrased as:

How much information is added by  $L$  to  $C$ ?

Knowing the organization of  $L$ , what could one say about  $C$ ?

# Resetting the problem

- Both linguistic and clinical data for the sample population can be organized in groups
- Let  $L$  and  $C$  be the names of those groups
- Questions in our Objectives could be rephrased as:

How much information is added by  $L$  to  $C$ ?

Knowing the organization of  $L$ , what could one say about  $C$ ?

# Resetting the problem

- Both linguistic and clinical data for the sample population can be organized in groups
- Let  $L$  and  $C$  be the names of those groups
- Questions in our Objectives could be rephrased as:

How much information is added by  $L$  to  $C$ ?

Knowing the organization of  $L$ , what could one say about  $C$ ?

# Entropy

## Entropy

$$H = - \sum_{k=1}^K P(k) \log_2 P(k)$$

- Entropy is a quantified expression of uncertainty of events (Shannon 1948), measured in bits
- The higher the entropy  $H$ , the most uncertain a given situation is
- If  $H = 0$ , we are certain of the outcome

# Entropy

## Entropy

$$H = - \sum_{k=1}^K P(k) \log_2 P(k)$$

- Entropy is a quantified expression of uncertainty of events (Shannon 1948), measured in bits
- The higher the entropy  $H$ , the most uncertain a given situation is
- If  $H = 0$ , we are certain of the outcome

# Entropy

## Entropy

$$H = - \sum_{k=1}^K P(k) \log_2 P(k)$$

- Entropy is a quantified expression of uncertainty of events (Shannon 1948), measured in bits
- The higher the entropy  $H$ , the most uncertain a given situation is
- If  $H = 0$ , we are certain of the outcome

# Entropy

## Entropy

$$H = - \sum_{k=1}^K P(k) \log_2 P(k)$$

- Entropy is a quantified expression of uncertainty of events (Shannon 1948), measured in bits
- The higher the entropy  $H$ , the most uncertain a given situation is
- If  $H = 0$ , we are certain of the outcome

# Conditional entropy

- Let L and C be two different ways of clusterizing K and K' groups
- The Conditional Entropy of C given L is expressed by the formula:

## Conditional entropy

$$H(C|L) = - \sum_{k=1}^K \sum_{k'=1}^{K'} P(k, k') \log_2 P(k'|k)$$

# Conditional entropy

- Let L and C be two different ways of clusterizing K and K' groups
- The Conditional Entropy of C given L is expressed by the formula:

## Conditional entropy

$$H(C|L) = - \sum_{k=1}^K \sum_{k'=1}^{K'} P(k, k') \log_2 P(k'|k)$$

# Conditional entropy

- Let L and C be two different ways of clusterizing K and K' groups
- The Conditional Entropy of C given L is expressed by the formula:

## Conditional entropy

$$H(C|L) = - \sum_{k=1}^K \sum_{k'=1}^{K'} P(k, k') \log_2 P(k'|k)$$

# Conditional entropy

- Informational contents in L can be expressed by  $H(L)$  bits; accordingly, informational contents in C are expressed in  $H(C)$  bits
- If someone knows  $H(L)$  bits of information, there would be still  $H(C|L)$  bits of uncertainty in the corpus
- $H(C|L) = 0$  would mean that we are certain of the outcome of C, because it is completely determined by L
- If  $H(C|L) = H(C)$ , L and C are independent group variables.

# Conditional entropy

- Informational contents in L can be expressed by  $H(L)$  bits; accordingly, informational contents in C are expressed in  $H(C)$  bits
- If someone knows  $H(L)$  bits of information, there would be still  $H(C|L)$  bits of uncertainty in the corpus
- $H(C|L) = 0$  would mean that we are certain of the outcome of C, because it is completely determined by L
- If  $H(C|L) = H(C)$ , L and C are independent group variables.

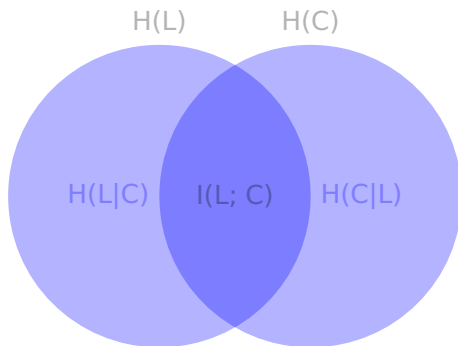
# Conditional entropy

- Informational contents in L can be expressed by  $H(L)$  bits; accordingly, informational contents in C are expressed in  $H(C)$  bits
- If someone knows  $H(L)$  bits of information, there would be still  $H(C|L)$  bits of uncertainty in the corpus
- $H(C|L) = 0$  would mean that we are certain of the outcome of C, because it is completely determined by L
- If  $H(C|L) = H(C)$ , L and C are independent group variables.

# Conditional entropy

- Informational contents in L can be expressed by  $H(L)$  bits; accordingly, informational contents in C are expressed in  $H(C)$  bits
- If someone knows  $H(L)$  bits of information, there would be still  $H(C|L)$  bits of uncertainty in the corpus
- $H(C|L) = 0$  would mean that we are certain of the outcome of C, because it is completely determined by L
- If  $H(C|L) = H(C)$ , L and C are independent group variables.

# Mutual information



- Reduction of uncertainty of one random variable due to knowing about another, or in other words, the amount of information one random variable contains about another [Manning & Schütze 1999]

# Variation of Information

## Variation of Information

$$VI(L, C) = H(L, C) + H(C, L)$$

- Evaluates how much information is to be gained when taking into account both linguistic and clinical clusterings

# Variation of Information

## Variation of Information

$$VI(L, C) = H(L, C) + H(C, L)$$

- Evaluates how much information is to be gained when taking into account both linguistic and clinical clusterings

# From Variation of Information

## Variation of Information

$$VI(L, C) = 0.582$$

- Since the quantitative measure of the variation of information stands reasonably far from 0, it indicates that there was a significant gain of information in taking into the linguistic clustering account

# From Variation of Information

## Variation of Information

$$VI(L, C) = 0.582$$

- Since the quantitative measure of the variation of information stands reasonably far from 0, it indicates that there was a significant gain of information in taking into the linguistic clustering account

# Dependency of clinical contents on linguistic contents

- Some MRT categories exhibit higher Mutual Information with Obit than others
- They also have higher entropies when compared to “trivial” categories
  - digestive apparatus:  $H = 2.92$
  - respiratory system:  $H = 1.00$

# Dependency of clinical contents on linguistic contents

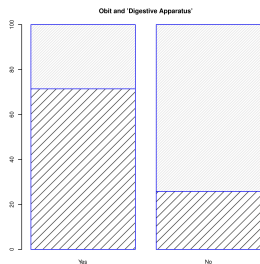
- Some MRT categories exhibit higher Mutual Information with Obit than others
- They also have higher entropies when compared to “trivial” categories
  - digestive apparatus:  $H = 2.92$
  - respiratory system:  $H = 1.00$

# Dependency of clinical contents on linguistic contents

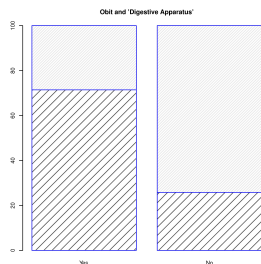
- Some MRT categories exhibit higher Mutual Information with Obit than others
- They also have higher entropies when compared to “trivial” categories
  - digestive apparatus:  $H = 2.92$
  - respiratory system:  $H = 1.00$

# Dependency of clinical contents on linguistic contents

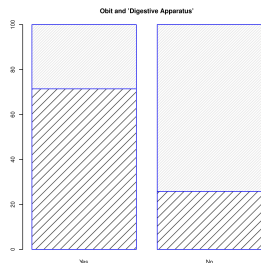
- Some MRT categories exhibit higher Mutual Information with Obit than others
- They also have higher entropies when compared to “trivial” categories
  - digestive apparatus:  $H = 2.92$
  - respiratory system:  $H = 1.00$



- Pearson's Qui-Square = 13.643;  $p = 0,000$
- Fischer's Exact Test:  $p = 0.0007041$
- Considering a significance level of 5%, we found dependency between Obits and MRT "Digestive Apparatus".



- Pearson's Qui-Square = 13.643;  $p = 0,000$
- Fischer's Exact Test:  $p = 0.0007041$
- Considering a significance level of 5%, we found dependency between Obits and MRT "Digestive Apparatus".



- Pearson's Qui-Square = 13.643;  $p = 0,000$
- Fischer's Exact Test:  $p = 0.0007041$
- Considering a significance level of 5%, we found dependency between Obits and MRT "Digestive Apparatus".

# Future research

- Test the same hypothesis over the remaining answers
- Provide the necessary means to make linguistic analysis partially or fully automatic
- Apply this methodology to other corpora of clinical interviews
- Compare our results to form-driven researches

# Future research

- Test the same hypothesis over the remaining answers
- Provide the necessary means to make linguistic analysis partially or fully automatic
- Apply this methodology to other corpora of clinical interviews
- Compare our results to form-driven researches

# Future research

- Test the same hypothesis over the remaining answers
- Provide the necessary means to make linguistic analysis partially or fully automatic
- Apply this methodology to other corpora of clinical interviews
- Compare our results to form-driven researches

# Future research

- Test the same hypothesis over the remaining answers
- Provide the necessary means to make linguistic analysis partially or fully automatic
- Apply this methodology to other corpora of clinical interviews
- Compare our results to form-driven researches