



# Can evaluation be application-independent?

Bente Maegaard  
Center for Language Technology  
University of Copenhagen



## The main question

Is it possible to evaluate technologies?

Or is it only possible to evaluate a technology for a particular application and use?

How do we define technology

How can we define application in a way that will be useful

We will start out by looking at very clear-cut examples, this will make it easier to understand



## Technology evaluation

We can probably safely say that one of the first – and very successful – technology evaluations were the speech recognition evaluation campaigns that were run in the US by DARPA

Speech recognition is a technology

It is used e.g. in dictation systems, dialogue systems etc.

Characterisation of speech recognition:

It is easy to measure the correctness. Either it is correct or not.

It cannot be more correct for some purposes than for others.

However, a recognition system which is generally good but has very big problems with proper names or with numbers, would not work very well in a telephone number service



## Application evaluation

### **Sparck Jones, K. & Galliers, J.R. (1996). Evaluating Natural Language Processing Systems:**

'Set up' or context of the application.

e.g. an MT system to be used for a certain purpose by a certain person with specific expertise

### **FEMTI - a Framework for the Evaluation of Machine Translation in ISLE**

(Hovy, King, et al.)

<http://www.issco.unige.ch:8080/cocoon/femti/st-home.html>



# FEMTI – evaluation purpose

## 1 Evaluation requirements

### 1.1 Purpose of evaluation

- 1.1.1 Internal evaluation
- 1.1.2 Diagnostic evaluation
- 1.1.3 Declarative evaluation
- 1.1.4 Operational evaluation
- 1.1.5 Usability evaluation
- 1.1.6 Feasibility evaluation
- 1.1.7 Requirements elicitation

### 1.2 Characteristics of the translation task

- 1.2.1 Assimilation
  - 1.2.1.1 Document routing or sorting
  - 1.2.1.2 Information extraction or summarization
  - 1.2.1.3 Search
- 1.2.2. Dissemination

Etc.



# FEMTI

## 2. System characteristics

### 2.1 Functionality

#### 2.1.1 Accuracy

### Correction rate 1

**Definition:** Correction rate defined as the ratio of the number of words corrected to the number of words in the translation (Van Slype)

**Method:** Count number of words corrected, number of words in initial translation.

**Measurement:** Ratio of number of words corrected to the number of words in the translation.



# FEMTI

## Correction rate 2

**Definition:** Correction rate defined as the number of insertions, deletions and substitutions - "edit distance" required to correct a text after translation (Ney and Niessen)

**Method:** Count the number of insertions, deletions, substitutions to correct a text. Note that this metric can be automated.

**Measurement:** Edit distance which is often a linear combination of the three counts.

### 2.1.1.1 Terminology

**Percentage of domain terms correctly translated.**

Etc.



## Examples of MT

An MT system which can only treat main clauses, no subordinate clauses – is this useful?

An MT system which can translate noun phrases only – is this useful?

In order for an evaluation to be useful, the purpose of the translation has to be known.



## Other technologies and applications: Can we make a FEMTI-like system?

FEMTI is a remarkable system for MT

Should be extended to other technologies and applications.

Need to establish the uses of the technologies and the possible metrics and methods.



## Resources for evaluation?

Irrespective of which evaluation method is used LRs will often be needed.

But the same LRs cannot be used again and again.

So, we should rather develop **methodologies and tools** for quick development of LRs for specific purposes.



## Acknowledgements

This presentation is inspired by a discussion in ELRA's  
Evaluation Committee

Present were:

Maghi King

Victoria Arranz

Khalid Choukri

Martine Garnier

Gregor Thurmair

And myself

