# All, and only, the typos

## LREC 2008, Marrakech, Morocco

Martin REYNAERT
Induction of Linguistic Knowledge
Tilburg University
The Netherlands
reynaert@uvt.nl

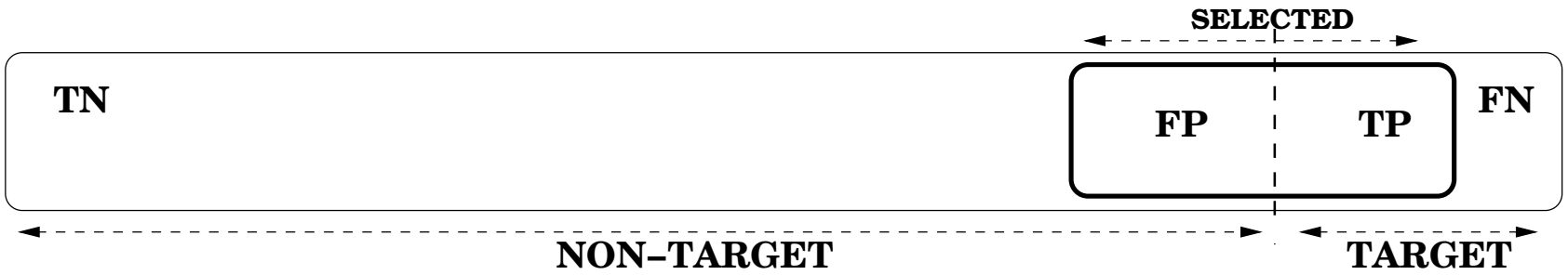
## Spelling and OCR-error correction evaluation

### Proposal for more complete and consistent evaluation

- Evaluation: goal is to see to what extent the task is successful: metrics used should be able to tell us that
- Accuracy measurements tell us only to what extent has been achieved
- Another goal of evaluation is to point the way forward towards perfect correction

# Spelling Correction: the TASK

Spelling correction = reduction of lexical variation caused by typos, OCR-errors, historical orthographical changes, ...



## EVALUATION: Measures

- TP = True Positives: real canonical forms for a particular error identified as such

- FN = False Negatives: real canonical forms for a particular error **not** identified as such

- FP = False Positives: correct words falsely reported to be variants for a particular correct word

- TN = True Negatives: correct words not reported to be variants for other correct words

# Evaluation: CONFUSION MATRIX

| | Target | Non-target |
|---|---|---|
| Selected | TP | FP |
| Not selected | FN | TN |

P = positive  N = negative

T = true  F = false

## Evaluation: METRICS

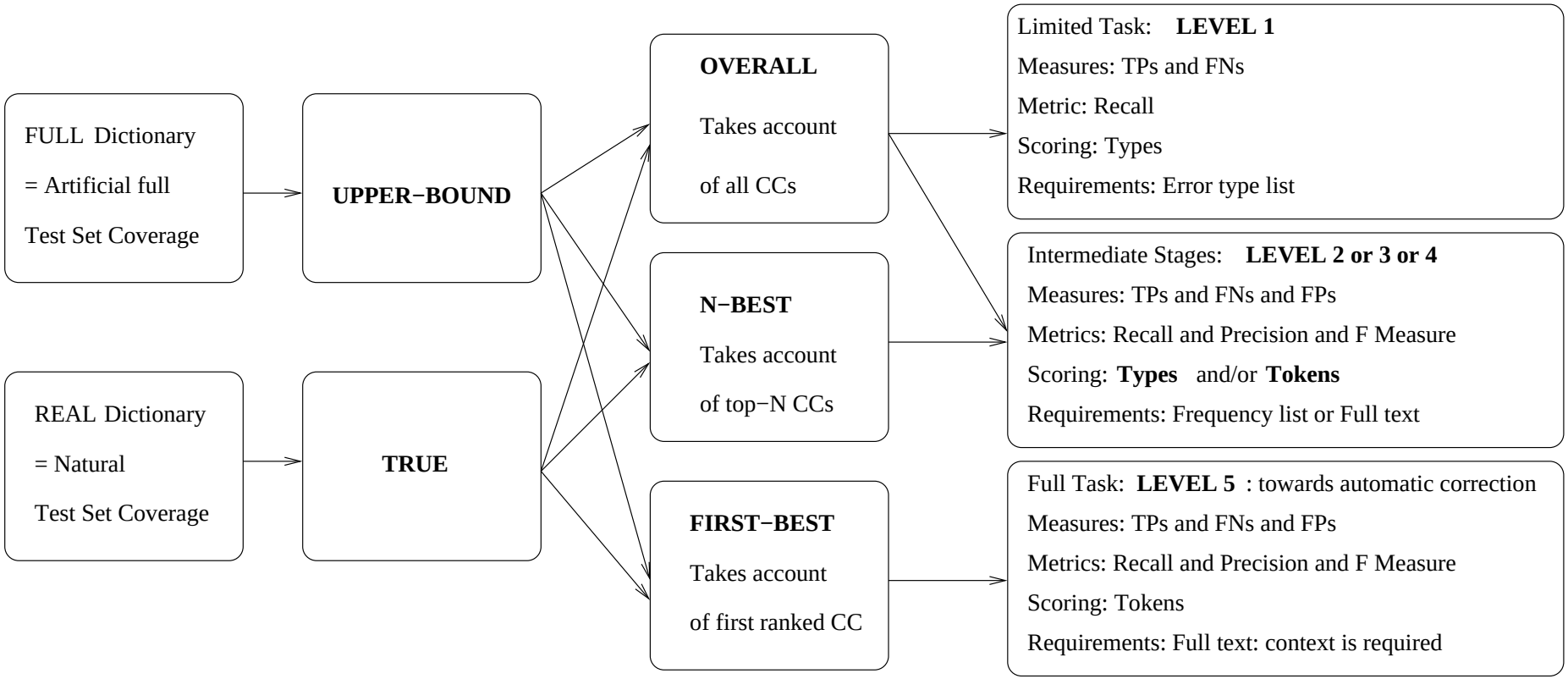
Van Rijsbergen (1975): From the TP, FN and FP we can derive recall and precision as follows:

$$\text{Recall} = \boldsymbol{R} = \frac{TP}{TP+FN} \qquad \text{Precision} = \boldsymbol{P} = \frac{TP}{TP+FP}$$

Since we deem recall and precision to be equally important, the harmonic mean of R and P, the simplified F measure, F, is given by:

$$\text{F-score} = \boldsymbol{F} = \frac{2\times R\times P}{R+P}$$

# Proposed Evaluation Framework



## Evaluation: STARTING POINT: the dictionary

**TWO OPTIONS:**

- Full dictionary coverage (= Artificial) $\Longrightarrow$ UPPER-BOUND

- Real dictionary (= Natural) $\Longrightarrow$ TRUE SCORES

## Evaluation: CORRECTION CANDIDATES

**THREE OPTIONS:**

- Overall: taking account of all CCs
- N-Best: taking account of top N ranked CCs
- First-Best: Taking account only of first ranked CC

## Evaluation: FIVE LEVELS

- All levels produce valid evaluations
- All levels focus on different aspects of the system
- The levels are complementary, results obtained are cumulative

- If 'claims to fame' are made: evaluation on lower levels only will not do!
- Lowest levels: more limited results, higher: more holistic view

## Evaluation: LEVEL 1: Core-correction mechanism

- How well is the algorithm capable of handling all the types of errors the system is said to be able to tackle?

- Measure the numbers of TPs and FNs.

- Metric: Recall

- Scoring: Types

- Test set: Error type list, paired with correct word forms

## Evaluation: LEVEL 2: Error detection

- What is erroneous and what is not? How many true and how many false alarms are raised?
- Measure the numbers of TPs, FNs and FPs.
- Recall, Precision $\Longrightarrow$ F Score
- Types and/or Tokens
- Test set: Frequency list or Full text: mix correct / incorrect word forms

## Evaluation: LEVEL 3: Suggesting correction candidates

- How often is the correct CC among the set of CCs?
- Measure the number of TPs in the set of CCs, those not present being FNs. FPs as measured on Level 2.
- Recall, Precision $\Longrightarrow$ F Score
- Types and/or Tokens
- Test set: Frequency list or Full text: mix

## Evaluation: LEVEL 4: N-best ranking

- How often is the correct CC among the n-best ranked CCs?

- (Likely smaller) number of TPs, the rest are the FNs. FPs as measured on Level 2.

- Recall, Precision $\Longrightarrow$ F Score

- Types and/or Tokens

- Test set: Frequency list or Full text: mix

## Evaluation: LEVEL 5: First-best ranking

- How often is the correct CC among the first-best ranked CCs? How often is the only CC the correct one?

- (Likely even smaller) number of TPs, the rest are the FNs. FPs as measured on Level 2.

- Recall, Precision $\Longrightarrow$ F Score

- Tokens

- Test set: Full text: context is required

## Clear and Concise Reporting

- 'We have conducted an Upper-bound, 5-best, Level 2 evaluation on types'

- 'We present a True, First-best, Level 5 evaluation on tokens using full text'

# Conclusion

**Framework proposed should allow for:**

- More complete evaluation
- More consistent evaluation
- More concise reporting

This work was undertaken within an NWO Exact Sciences Hefboom-project