

# **A Semantic-oriented Readability Checker for German**

**Tim vor der Brück, Sven Hartrumpf**

<http://pi7.fernuni-hagen.de/>

**FernUniversität in Hagen**

**58084 Hagen, Germany**

**2007-10-06, Poznań**

# Overview

<b>1</b>	<b>Readability - Definition</b>	<b>2</b>
<b>2</b>	<b>Readability Indicators</b>	<b>3</b>
<b>3</b>	<b>DeLite</b>	<b>4</b>
<b>4</b>	<b>Syntactico-semantic analysis</b>	<b>5</b>
<b>5</b>	<b>Calculation of the global readability score</b>	<b>7</b>
<b>6</b>	<b>Semantic Indicators</b>	<b>13</b>
<b>7</b>	<b>Evaluation</b>	<b>18</b>
<b>8</b>	<b>Graphical user Interface</b>	<b>19</b>
<b>9</b>	<b>Conclusion</b>	<b>21</b>

# 1 Readability - Definition

- ◇ Definition: Edgar Dale and Jeanne Chall (1949): The sum total (including all the interactions) of all those elements within a given piece of printed material that affect the success a group of readers have with it. The success is the extent to which they understand it, read it at an optimal speed, and find it interesting.
- ◇ We focus only on understandability
- ◇ Typical readability function:  
 $f : A^* \rightarrow [0, 1]$  with  
$$f(t) = w_1 v_1(t) + \dots + w_n v_n(t)$$
- ◇  $v_1, \dots, v_n$  correspond to readability indicator values
- ◇  $w_1, \dots, w_n$  parameters, often determined by linear regression

## 2 Readability Indicators

Often used indicators in traditional formulas:

- ◇ Average sentence length
- ◇ Average word length
- ◇ Word frequency analysis

Drawbacks of current readability functions:

- ◇ Current readability checkers only use surface-oriented indicators
- ◇ Thus: Only rough approximation of cognitive difficulties possible

Our indicators: based on a a deep syntactico-semantic analysis

### 3 DeLite

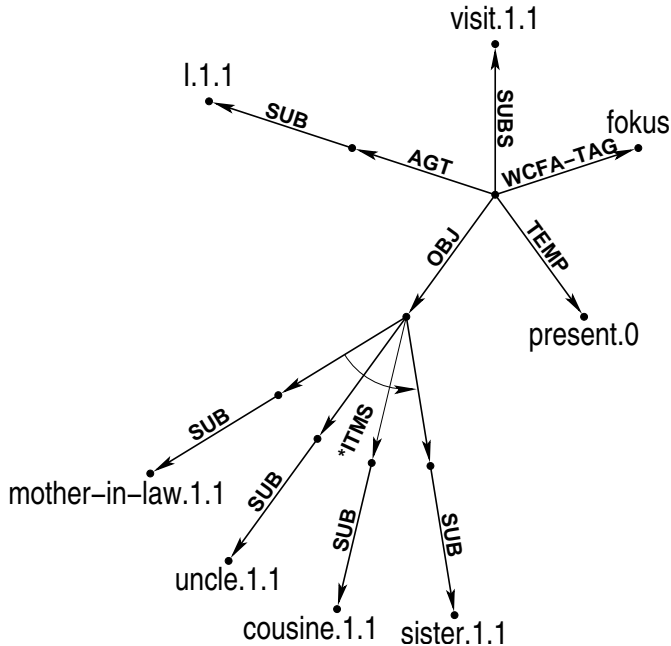
- ◇ Uses semantic readability indicators in addition
- ◇ Semantic readability indicators based on a semantic network in the Multi-Net formalism
- ◇ Functionality
  - Highlights difficult to read text passages
  - Calculates a global readability score

## 4 Syntactico-semantic analysis

### Overview

- ◇ Deep syntactico-semantic analysis
- ◇ Exploits a semantic-oriented lexicon
- ◇ Sentences are rejected which
  - violate syntactic constraints (Example: Pete not goes to school.)
  - violate semantic constraints (Example: The pie is eating the apple.)

## Example of a semantic network



I visit the  
mother-in-law, the uncle,  
the cousin, and the sister.

- ◇ AGT: Agent/actor in the sentence, here *I*
- ◇ SUBS: Action, here *visit*
- ◇ OBJ: The persons visited, combined by the function ITMS

## 5 Calculation of the global readability score

Dr. Peters invites  
Mr. Müller for dinner.  
It's his birthday today.

- Indicator operating on word level
- Indicator operating on sentence level

Analyze the given input text

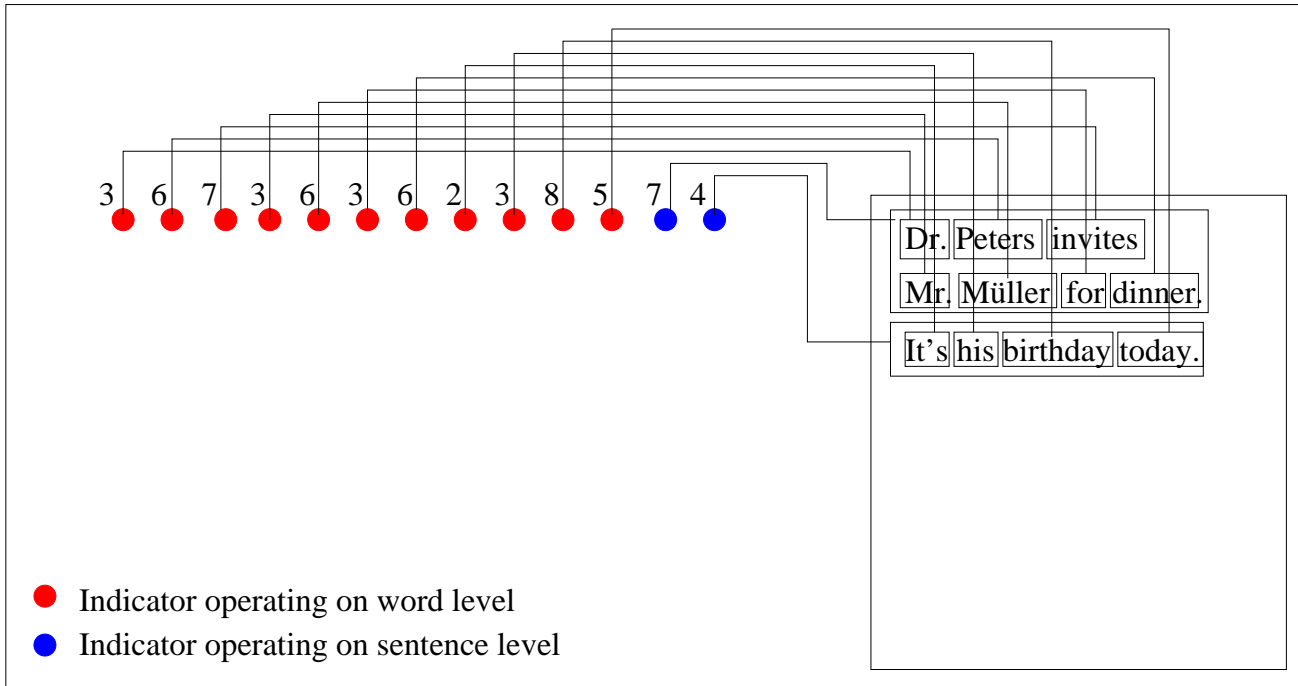
# Calculation of the global readability score (contd.)

Dr.	Peters	invites	
Mr.	Müller	for	dinner,
It's	his	birthday	today.

- Indicator operating on word level
- Indicator operating on sentence level

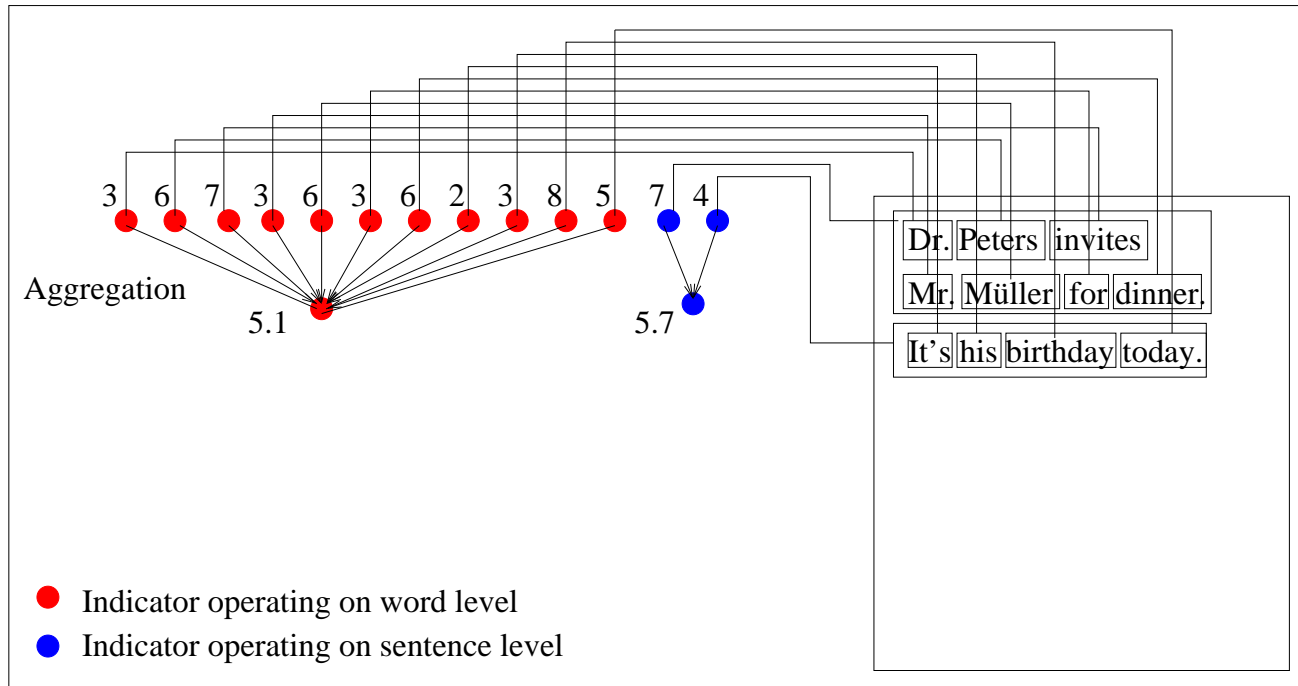
1st step: Determination of document structure

# Calculation of the global readability score (contd.)



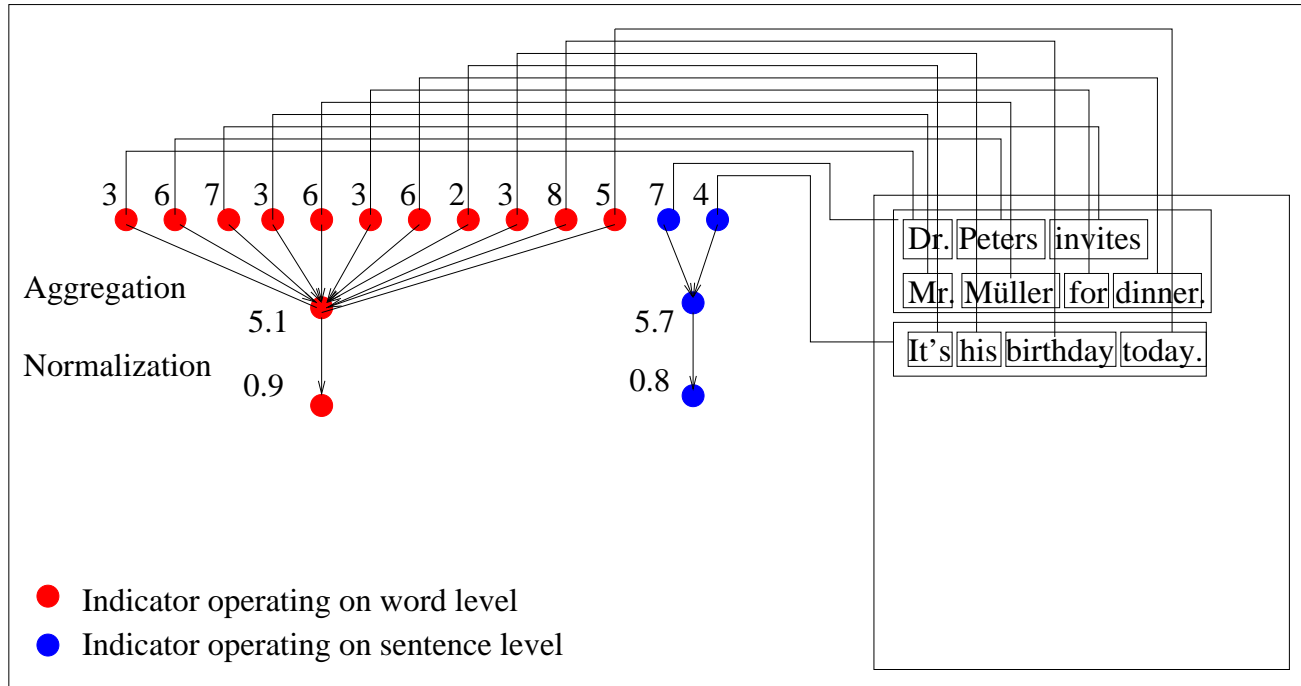
2n step: Calculation of indicator values

# Calculation of the global readability score (contd.)



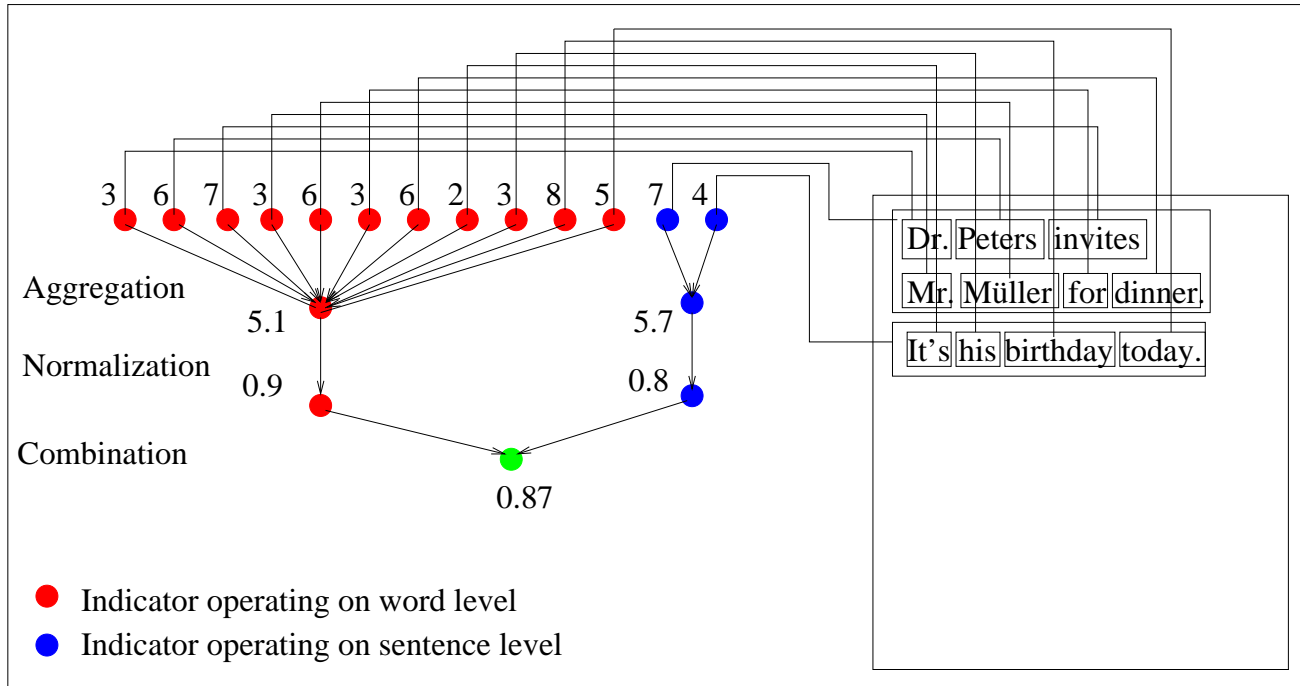
3rd step: Aggregate indicator values

# Calculation of the global readability score (contd.)



4th step: Normalize indicator values

# Calculation of the global readability score (contd.)



Final Step: Combine indicator values

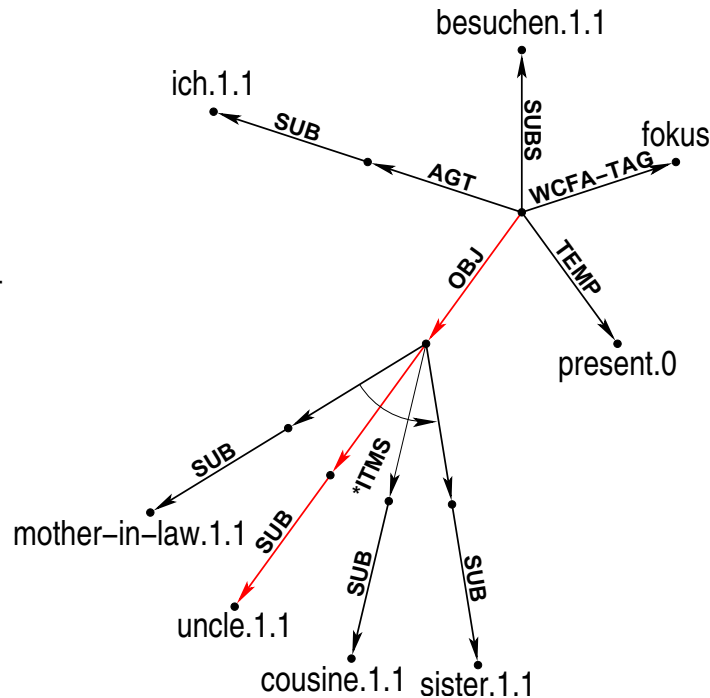
# 6 Semantic Indicators

## 6.1 Longest Path

- ◇ Long pathes in the semantic network correspond often to sentences with complicated dependency structures
- ◇ Thus: An indicator Longest Path is used

## Longest Path (contd.)

Example Sentence “I visit the mother-in-law, the uncle, the cousin, and the sister.”

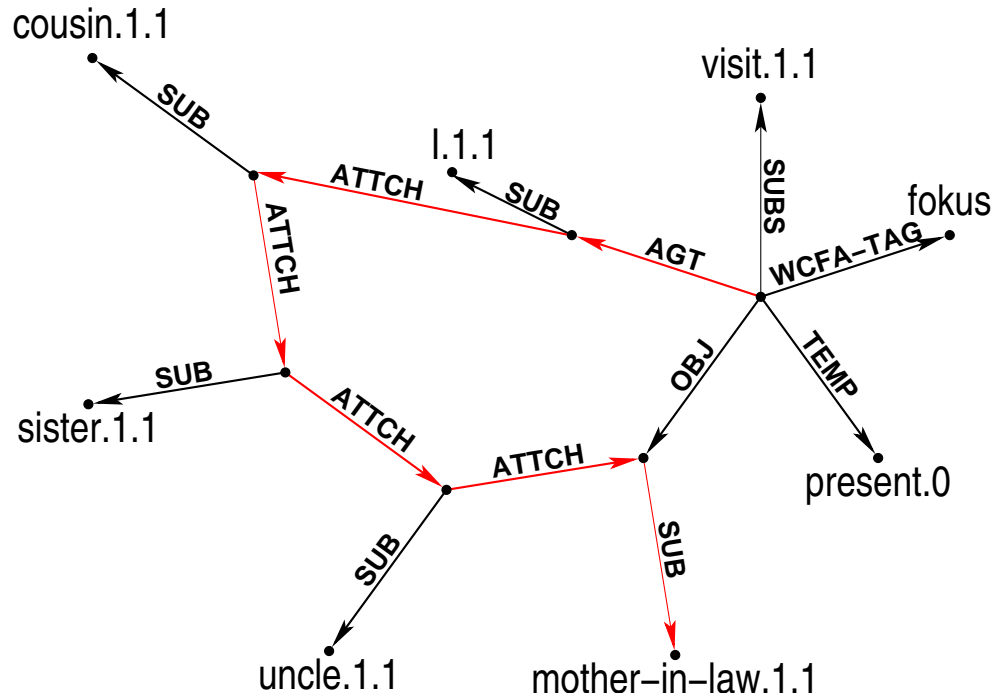


- ◇ Parallel interpretation possible
- ◇ Longest path: 3

## Longest Path (contd.)

Example Sentence “I visit the mother-in-law of the uncle of my cousin’s sister.”

- ◇ Sequential interpretation necessary
- ◇ Longest path: 6



# Anaphers

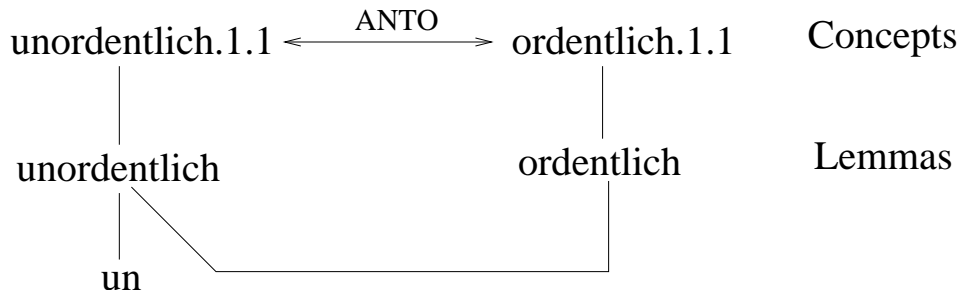
## Coreference Resolution

- ◇ Example Sentence: *Dr. Peters invites Mr. Müller for dinner.* because he has birthday today.
- ◇ The pronoun he can either relate to Dr. Peters or Mr. Müller
- ◇ Number of reference candidates: 2 for *he*
- ◇ Other coreference-based indicators: *No antecedent found* and *distance between pronoun and antecedent in words and sentences*

## Negations

Indicator *Negations* recognizes:

- ◇ Negative expressions like never, nowhere
- ◇ Several types of double negations
- ◇ Negation prefixes



## 7 Evaluation

Total weight of semantic indicator are 10%  
(30% including semantic network quality)

Quality of the SN	0.360
Passive with AGENT	0.209
Pronoun Reference Distance	0.203
Number of Propositions per Sentence	0.201
(Double) Negations	0.189
Connectivity of Discourse Entities	0.186
Longest Path	0.108

# 8 Graphical user Interface

Peter talks with John. He still goes to school.

**Readability indicators** many reference candidates for a pronoun

- Understandability index (Amstad):122.35
- Number of words:10
- Number of syllables:9
- Number of sentences:2
- Average sentence length:5
- Type-token-ratio (lemmata):0.9
- Type-token-ratio (word forms):0.9
- Number of abbreviations:0
- Number of acronyms:0
  
- Morphological level:97.47%
- Lexical level:50.82%
- Syntactic level:93.67%
- Semantic level:96.09%
- Discourse level:80.84%


[XML report R1 \(text structure\)](#)

[XML report R2 \(indicators\)](#)

[XML report R3 \(scores and weights\)](#)

**Enter the text to be analyzed (copy & paste):**

Peter talks with John. He still goes to school.

Readability score:  
  
(82/100)

**Morphological level** ⓘ  
No problems found

**Lexical level** ⓘ  
Word frequency  
Lexical ambiguity

**Syntactic level** ⓘ  
No problems found

**Semantic level** ⓘ  
No problems found

**Discourse level** ⓘ  
Reference ambiguity

- ◇ difficult-to-read text passages are highlighted
- ◇ displays a global readability-score

## 9 Conclusion

### Semantic readability indicators

- ◇ are in some cases superior to surface type indicators
- ◇ have an influence of 10% in our readability formula
- ◇ their influence is expected to be higher when coverage is improved
- ◇ are expected to be important for future readability formulas