Factors of Readability of Polish Texts: A Psycholinguistic Study

Edyta Charzyńska
Warsaw School of Social Sciences and Humanities, Poland

Łukasz Dębowski
Institute of Computer Science, Polish Academy of Science, Warsaw, Poland

QUALICO 2014: Olomouc, 29th May - 1st June 2014

This research was supported by the National Science Centre under Grant 2011/03/BHS2/05799.
TEXT READABILITY

1. Legibility of the printed material as well as its layout or typography.
2. Ease of reading due to the interest-value or the aesthetics of writing.
3. Ease of comprehension due to the style of writing (Klare, 1963; Samson, 1993).
HOW CAN WE MEASURE TEXT READABILITY?

Measurement methods

- reading comprehension tests
- analitical methods: readability formulas
- psycholinguistic method: cloze test
- other methods: rate of reading, oral reading errors, expert/reader judgments of difficulty
READABILITY FORMULAS

• The majority of readability formulas are based on **TWO FACTORS** (Chall & Dale 1995):

  - **Lexical or semantic features**
    - word length,
    - word familiarity,
    - word frequency
  - **Sentence or syntactic complexity**
    - sentence length – the average number of words per sentence
READABILITY FORMULAS FOR SLAVIC LANGUAGES

• The first readability formula for Slovak language, constructed by Józef Mistrík (1968), included 3 factors:

✓ $\lambda_s$ – average sentence length (in words),
✓ $\lambda_v$ – average word length (in syllables),
✓ $I$ – inverse of type/token ratio

\[ R = 50 - \frac{\lambda_s \cdot \lambda_v}{I} \]
READABILITY FORMULA FOR POLISH

• Walery Pisarek (1969; 2007)

\[ T = \frac{\sqrt{T_s^2 + T_w^2}}{2} \]

- \( T \) - text difficulty
- \( T_s \) - average sentence length (in words)
- \( T_w \) - percent of „difficult” words
  (words with four or more syllables in their base form)

The formula has never been empirically tested.
CLOZE PROCEDURE

• Developed by Wilson Taylor in 1953;
• Cloze – *clozure* in Gestalt psychology;
• Readers are given a passage from which some words have been deleted;
• Their task is to guess the deleted words;
• Validity of the cloze method is generally accepted (Rankin, 1959; Bormuth, 1967, 1968).
The purpose was to examine validity of the readability formula for Polish developed by Pisarek (1969);

N = 1,309 Polish native speakers;

Sex: 844 (64.5%) women and 464 men (35.5%), 1 missing data;

Age: $M = 34.87$ ys; $SD = 16.33$; min. = 15; max. = 83
Educational structure of the research sample

<table>
<thead>
<tr>
<th>EDUCATION</th>
<th>NUMBER OF PEOPLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>elementary</td>
<td>85</td>
</tr>
<tr>
<td>lower secondary</td>
<td>235</td>
</tr>
<tr>
<td>vocational</td>
<td>90</td>
</tr>
<tr>
<td>secondary</td>
<td>413</td>
</tr>
<tr>
<td>higher</td>
<td>486</td>
</tr>
</tbody>
</table>
METHODS

• a sample of 15 differentiated texts, each consisting of about 300 words:
  ✓ 2 scientific texts,
  ✓ 2 texts from the secondary school handbooks,
  ✓ 2 enactments,
  ✓ 2 official letters,
  ✓ 2 instructions,
  ✓ 2 law brochures,
  ✓ 3 journalistic articles from various fields.

• The readability of each text was measured by 3 different tests: multiple-choice (4 questions with 4 options), cloze (50 gaps) and open-ended questions (5 questions).
PROCEDURE

• Each person was randomly given 3 different texts, followed by the multiple-choice, cloze or open-ended question test;

• Participants filled in the tests in the presence of the member of the experimental team, usually a student;

• The task completion took 30-40 minutes on average.
ANALYSIS

• Statistical descriptives for each text;
• Pearson correlation coefficient between the comprehension tests and the two linguistic variables from Pisarek's formula;
• Pearson correlation coefficient between the comprehension tests and the full formula of Pisarek;
RESULTS

- Correlations between comprehension tests and average sentence length

\[ r_{\text{Cloze}} = -0.607 \ (p = 0.016) \]
\[ r_{\text{Open}} = -0.401 \ (p > 0.05) \]
\[ r_{\text{MC}} = -0.337 \ (p > 0.05) \]
RESULTS

• Correlations between comprehension tests and word difficulty

\[ r_{\text{Cloze}} = -0.276 \ (p > 0.05) \]
\[ r_{\text{Open}} = -0.477 \ (p = 0.072) \]
\[ r_{\text{MC}} = -0.597 \ (p = 0.019) \]
RESULTS – correlation between Pisarek's formula and the cloze test

\[ r_{\text{Cloze}} = -0.656 \ (p = 0.008) \]
RESULTS – correlation between Pisarek's formula and the open-ended questions

$r_{\text{Open}}: -0.519 \ (p = 0.047)$
RESULTS – correlation between Pisarek's formula and the multiple-choice test

\[ r_{MC} = -0.539 \quad (p = 0.038) \]
DISCUSSION

• Validity of Pisarek’s formula was partially confirmed; correlations between the formula and comprehension tests were significant, but not high (r in range between -0.52 and -0.66);

• How to improve the formula?

  Modify the equation by adding coefficients?

  Exclude words with 4 and more syllables which occur frequently in Polish texts?

  Include new variables: abstract nouns? formal terms?
Thank you for your attention!

echarzynska@swps.edu.pl
ldebowsk@ipipan.waw.pl