Index of Readability

Giuseppe Di Modica
Course of
Processing of Natural Language

A.A. 2003/2004

Legibility: the problem

Over the past decades, the issue *of readability* has gained increasing attention However, the assessment of the readability of a text does not always take account of all linguistic factors which may hinder or impede its understanding.

Legibility: the problem

The *legibility of a* text, in other words, is often assessed by the expression of subjective judgments inherent *in the* simplicity of the same
greater or lesser presence of " *difficult words* "
that do not always coincide with its *comprehensibility. For* legibility, in *reality, we must mean the system* of a text that makes it more or less clear and understandable on the basis of a very large number of linguistic characteristics in combination, regardless of the complexity of the topics contained.
Of course, however, the easier a text is to read, the more likely it is to be easy to understand!

Study results

Since the 1980s, numerous studies have been carried out which have determined that the factors of legibility of a text are 3:

1. the graphic aspect:
   presence of images, tables and drawings;
   organization of the text in chapters, paragraphs and sub-paragraphs and titration of these partitions;
   the presence of special characters to signal definitions and vocabulary;

Study results

2. the length of the sentences:
   - the longer a sentence is, and therefore full of subordinations, the less easy and immediate it will be;
3. the length of the words within each individual sentence:
   - the longer a word is, the more information it transmits;
   - the presence of many long words can make a sentence too dense in meaning and therefore not easy to read and understand.

Readability on the web

It is gaining increasing importance, due to the development of the internet not only as a means of communication but also as a privileged source of collection and dissemination of documents.

These are not only advantages...

Physical readability Linguistic readability

Physical readability

If we reproduce a paper document on the web very hardly this will be equally readable

- for the simple reason that the computer monitor is not very suitable for reading for "physical" reasons

Being linked to a projection of light beams on a screen, reading on the web decreases the ability to concentrate the reader, and quickly becomes tiring

- 25%-30% slower than on paper

Physical readability

Result: the web user rarely reads word for word, preferring rather "scan" the page in search of "landing" points that attract his attention

(bold, titles, spacing, etc...)

and quickly jumping from one side of the page to the other.

Linguistic readability

Linguistic readability, on the other hand, concerns the use of language in all its components:

- choice of terms
- choice of syntax used content
- articulation

The problem of linguistic readability also mainly concerns texts with an informative purpose.

We don't read *I Promessi* Sposisolo for information, but for the taste we get.

It's not the same for a drug liar or a tax return form, from which we expect precise and understandable instructions.

We should not be surprised, however, if the legibility of a page by Manzoni and that of most of the liars are equal.

Readability index

The problem therefore arises of establishing *standards of* document readability

Objective centered thanks to the creation of readability indexes

There are several, four main ones:

- Flesch index
- Index of Kincaid
- Index of Gunning's Fog
- GULPEASE index

Flesch index

The first to declare that the readability of a text is a measurable concept, was the American scholar Rudolph Flesch.

He was also the first to propose a method of measuring it, known as the Flesch index.

According to the studies of Flesch

- exhibited in his book *The Art of Plain Talk* (1946)

a text is difficult when:

- contains many subordinates
  - syntactic difficulty
- and many abstract words
  - semantic difficulty

Flesch index

The Flesch formula calculates readability by taking into account the average length of words, measured in syllables, and the average length of sentences, measured in words.

In practice:

- a long word is generally used less frequently than a short one;
- a long sentence is usually more complex, from the point of view of syntax, than a short one.

Flesch index

The 2 language variables of the Flesch index:

- the average length of spreadsheet words in syllables per word (S),
- and the average length of the frasiespressa in words per sentence (W).

La Formula di Flesch per la lingua inglese

Ease of reading =

206,835 - 0,864 S - 1,015 W

S = sillabe di 100 parole, W = media di parole per frase

Disadvantages...

The *formula of Flesch*
which owes its diffusion precisely to its simplicity
has two drawbacks:

1. the formula has been designed for English and is, therefore, calibrated on the morphological and syllabic structure of this language;
2. the problem of counting syllables.

In fact, precisely this type of calculation is particularly complex within the Italian language, since it is not completely formalizable by means of rules of general scope, except by using statistical estimates, the limit of which, unfortunately, is that of not being able to describe and reproduce exactly the hyphenation of the single words of a text.

Adaptation of the formula

The fact that the formula was born for English was addressed by Roberto Vacca, who, in 1972, adapted the parameters of the formula to the Italian language.

The index was named Flesch-Vacca

La Formula di Flesch per la lingua italiana adattata da Franchina - Vacca (vers 1972)

Facilità di Lettura=

206 - 0.65 S - W

S = sillabe di 100 parole, W = media di parole per frase

Index of Flesch-Vacca

La Formula di Flesch per la lingua inglese

Ease of reading =

206,835 - 0,864 S - 1,015 W

S – sillabe di 100 parole, W – medie di parole per frase

La Formula di Flesch per la lingua italiana adattata da Franchina - Vacca (vers 1972)

Facilità di Lettura=

206 - 0.65 S - W

S = sillabe di 100 parole, W = media di parole per frase

Once again:

W average number of words per sentence in a sample of one hundred words
S number of syllables per 100 words
206 is the constant applied to maintain values between 0 and 100

before it was 206,835

0.65 is the constant referring to the average length of Italian words
before it was 0.864

Index of Flesch-Vacca

The results of the formula fluctuate between 0 and 100

"0" means the lowest readability,

"100" means the highest readability.

Because of its ease of use, the index can be used to analyze both short and long texts. In the latter case it is necessary to operate on a sample of the text.

Index of Kincaid

It's a modified version of Flesch's original formula.

This index also calculates complexity based on the average number of syllables per word (S) and the average number of words per sentence (W).

The result obtained is an approximate measure of the number of school years

Education level (American though!)

that the reader should have done to understand the content of the text.

Index of Kincaid

Education = (0.39 x W) + (11.8 x S) - 15.59

Results between 0

negative numbers are considered as 0

and 12

numbers over 12 are reported as 12

Values between 6 and 10 indicate that text can be read easily by most people.

For "technical" documents it is good to have values > 10

Age of reader = Education + 5

Index of Gunning's Fog

This index is similar to the previous one and reflects, in an approximate way, the minimum number of school years that a person must have attended in order to easily read the text under examination.

The formula is:

0.4 x (average number of words in a sentence + percentage of difficult words in the text)

where words are considered "*difficult*" if they have more than two syllables, except for a few words of three syllables with certain final syllables.

Results over 17 are reported as 17, where 17 is considered the postgraduate level

Index of Gunning's Fog

If we apply the index of Gunning's Fog we discover that...

Fog Index	Resources
6	TV guides, The Bible, Mark Twain
8	Reader's Digest
8 - 10	Most popular novels
10	Time, Newsweek
11	Wall Street Journal
14	The Times, The Guardian
15 - 20	Academic papers
Over 20	Only government sites can get away with this, because you can't ignore them.
Over 30	The government is covering something up

Considerations

What the first three indices have in common
- if we neglect for a moment that of Flesch-Cow

is to have been made on the (and for the) English language.

The Flesch-Vacca index itself is, however, an approximation of the original to fit the Italian language

The need for a legibility index created specifically for our language arises.

GULPEASE index

In 1987 a group of linguists from La Sapienza University in Rome gathered around Tullio de Mauro to form the GULP
- University Linguistic Pedagogical Group

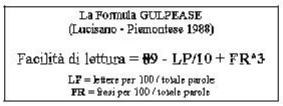
The GULPEASE index is born

To formulate it, scholars have verified in different types of reader the real comprehensibility of a *corpus of texts*.

Among the factors considered in the formula are the number of letters per word and the number of words per sentence.

GULPEASE index

GULPEASE index: the formula



The values obtained are, as for the Flesch index, included in a scale ranging from 0 to 100

We'll see better soon...

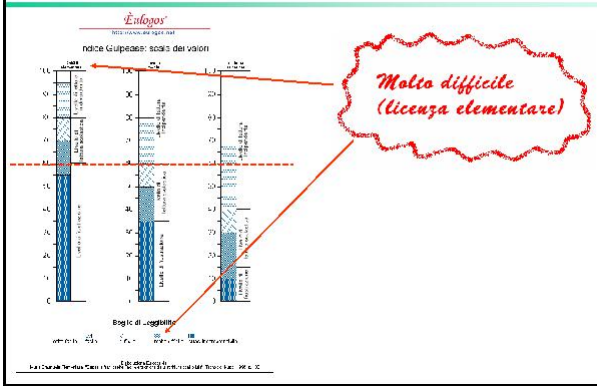
GULPEASE index

Readers with elementary education can easily read texts with an index above 80.

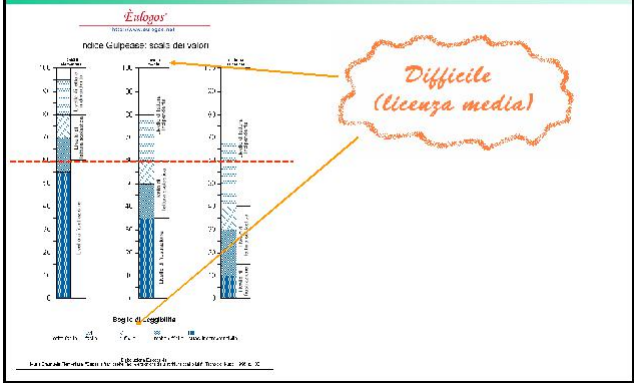
Readers with an average education can easily read texts with an index above 60.

Readers with higher education can easily read texts with an index above 40.

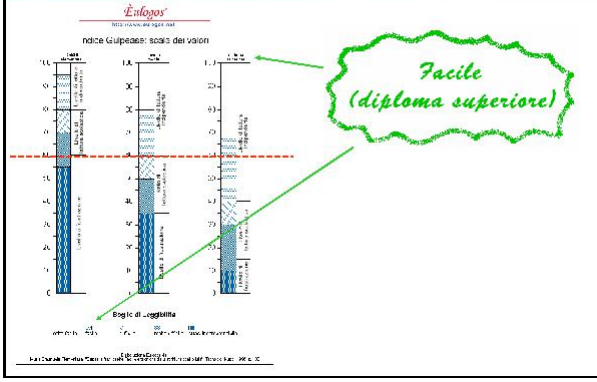
GULPEASE index



GULPEASE index



GULPEASE index



GULPEASE index: the advantages...

The most obvious is the first readability formula calibrated directly on the Italian language

No less important is that it has the advantage of calculating the length of the words the letters, and no longer in syllables.

It lends itself, therefore, well to be automated

But is readability used?

Quite...

In the United States the problem of the readability of written texts has been in evidence for more than a century.

Not only. It was considered so important that 29 states signed the "Plan Language Act".

The "Law on Speaking Clearly" requires companies to write in clear and understandable language any kind of message. In addition, this legislation requires the Flesch index to be used as a legibility criterion. All public documents must have a legibility index of not less than 45.

What about in Italy?

As for the Italian situation, a judgment (364 of 23 March 1988) of the Constitutional Court admits that the citizen can ignore the law if formulated in an incomprehensible way.

But more should be done, in fact...

Analysis of readability

Analyzing

with the index of Flesch-Cow

some documents we realize that:

first-grade reading books have an index around 100;

some documents of the bureaucracy reach even the -100;

Moravia was found to have an average readability index of 80;

First Levi at 65.

But not only that...

Analysis of readability

Returning to Manzoni's example, it is not surprising to know that the GULPEASE index of the first page of *"The Betrothed"*

"That branch of Lake Como..."

is 49 and that the Flesch-Cow is of 21. It is worrisome, however, to verify that an official text has index GULPEASE 43 and

12 of the Flesch-Cow index.

And this is the document on the issue of international debt published on the website of the Ministry of Foreign Affairs.

And on the web?

Things are not so much better... A recent survey

conducted by Genesio

which used a sample of some of the most visited information sites in Italy, wanted to verify readability on the web.

And the results haven't been very encouraging...

Readability on the web

The sites that have been examined belong to several areas:

- formation,
- official information,
- journalism,
- business communication.

The factors taken into account were:

- Gulpease index,
- Flesch-Vacca index,
- Gunning's Fog index,
- Kincaid index,
- total words, words per sentence, letters per word.

Survey of Genesio

As regards indices, the most relevant research items are the Gulpease index and the Flesch-Vacca index.

The minimum readability threshold for the Gulpease index was chosen between 50 and 60, for the Flesch-Vacca index 45

- that is the same parameter used by American law in this regard.

The results were mostly mixed. Worrying and reassuring at the same time. Let's see them a little better in detail...

Genesio survey results

The lowest readability data remain those of public sites.

While it is the same public service that has promoted the campaign for readability, government sites insist on using a language model that is well below the readability index.

Genesio survey results

The site of the Foreign Ministry, for example, has a Gulpease index of 41.7 and Flesch-Cow index of 23.

Genesio survey results

Only slightly higher are the values of the Government site (Gulpease 45.7, Flesch 36.7)

Genesio survey results

In addition to public sites, even those of the sample belonging to large companies show very low readability indices.

One example is Microsoft. And to think that this company leader in the field of informatics has shown particular sensitivity to the problems of readability, including among the operational tools of Word the statistics of readability.