Comparative Evaluation of Template Systems: Metric Definitions

Authors omitted
for double-blind review

I. Introduction

This document summarizes detailed metric definitions for the *Comparative Evaluation of Template Systems* study. In addition, details on the comparison of guidelines underlying some of these metrics are provided.

II. Requirement Phrasing Guideline Comparison

Table I summarizes the six examined phrasing guidelines and their similarities in rules. The references within the cells indicating that rules are covered by some guideline point to the respective rule identifier within the original source document or the respective section where no identifiers are provided by the guideline.

It can be seen from Table I that the examined guidelines have different focus. None of them contains all 39 rules. The INCOSE guide [1] covers most of the aspects with 30 rules, directly followed by the SOPHIST rules [2], covering 26 rules. The lowest number of rules is covered by ECSS-E-10-06C [3] (12 rules) and drafting rules [4] (13 rules), while ISO [5] (17 rules) and NASA [6] (19 rules) guidelines cover slightly more.

Only four rules are contained in all six guidelines: (6)[1] “use simple sentence structure”, (17) “avoid vague terms”, (29) “use context free phrasing”, and (36) “express one atomic need”.

Five further rules are contained in all but one guideline: (8) “use active voice” is only missing in ECSS-E-ST-10-06C. Rules (25) “separate rationale from sentence” and (33) “use solution free phrasing” are only missing in the ECSS drafting rules. This is a bit surprising, as most ECSS standards appear to follow these rules nevertheless. Rules (5) “use defined modal verb for liability” and (7) “use appropriate abstraction level” are solely absent in the SOPHIST rules. This is astonishing, because both rules are prominently contained in related work by SOPHIST [7–9]. In particular, (5) is emphasized for MASTeR templates [7, 10, 11], which aim to incorporate SOPHIST rules, and is measured by Wolf and Strößner’s [9] *Classifiable* metric.

ECSS-E-ST-10-06C is a subset of the INCOSE guideline focusing on those rules especially relevant to unambiguity. Similar, the NASA guideline seems to be oriented along the INCOSE guide, but with three exceptions only also covered by the SOPHIST rules.

Generally, the SOPHIST rules and INCOSE guidelines are the only ones with unique features. While SOPHIST appears to focus on linguistic effects, INCOSE focuses more on the reduction of complex syntactic structures. The union of both covers all 39 rules and subsumes the other four rule sets.

III. Metric definitions

Metrics are documented in Table II-VIII, following the template suggested in IEEE 1061 [12] under omission of some attributes not relevant in the context of this evaluation, namely, *costs*, *benefits*, *impact*, *training required*, and *validation history*. For conciseness, several metrics are aggregated in one table based on commonalities in calculation.

[1]numbers refer to identifiers in Table I

Phrasing Rules	Phrasing Guidelines for Requirements	[5] ISO/IEC/IEEE 29148	[2] SOPHIST Rules	[1] INCOSE Guide	[3] ECSS ST-E-10-06	[4] ECSS Drafting Rules	[6] NASA Guide
1	use only one sentence		R4+9	R11+18	8.3.1		
2	avoid unnecessary words		R15			5.2.3	C.4
3	use only one process verb		R4-5+15	R2			C.4
4	avoid extensive punctuations		R14-15	R14			
5	use defined modal verb for liability	5.2.4		R1	8.3.2	5.2.1	C.1
6	use simple structured sentence (full sentence with noun and verb, no flowery phrase or verbiage ...)	5.2.4	R14-15	R2+41	8.3.1	5.2.3	C.2+4
7	use appropriate abstraction level	5.2.5		R3+31	8.3.1	5.2.2	C.4
8	use active voice	5.2.4	R1	R2		5.2.5	C.2
9	use precise verb		R2+15				
10	avoid nominalization		R3				
11	avoid light verb construction		R4				
12	use full verb		R1-5				
13	avoid comparison	5.2.7	R8				
14	use clear comparison		R8				
15	use definite articles		R10-11	R5			
16	use defined units			R6		5.3.2.2	
17	avoid vague terms	5.2.4+7	R2+8+12+15	R7	8.3.3	5.2.3+C	C.4
18	avoid escape clauses	5.2.7		R8			
19	avoid open ended clauses	5.2.7		R9			
20	avoid superfluous infinitives			R10			
21	use correct grammar + spelling			R12-14			C.3
22	avoid negations	5.2.4		R16	8.3.1		C.3
23	avoid /			R17			
24	avoid combinators		R9	R19	7.2.3+8.2.7		
25	separate rationale from sentence	5.2.5+7	R14	R20	8.2.2		C.3+4
26	avoid parentheses			R21			
27	avoid group-nouns		R10-12	R22			
28	avoid pronouns	5.2.7		R24			C.4
29	use context free phrasing	5.2.7	R6-7+16-18	R23+25	8.2.8	5.2.3	C.4
30	avoid absolutes	5.2.7		R26			
31	use explicit conditions	5.2.4	R11+16-18	R27+35			
32	use clear condition combinations		R16-18	R15+28			
33	use solution free phrasing	5.2.4+7	R13	R31	4.1+8.3.1		C.2+4
34	use clear quantifiers		R8+10-11	R32+34			C.2
35	use value tolerances			R33	8.2.10	5.3.2.3	C.2+4
36	express one atomic need	5.2.5	R9+15	R11+18	7.2.3+8.2.7	5.2.3	C.4
37	use clear preconditions		R13+16-18	R35		C.3.2.2	C.3
38	use clear business logic	5.2.4	R6+13			5.2.3	C.4
39	use clear subject		R6			5.2.5	C.2+4

Table I: Requirement Phrasing Guidelines and their Rules

Figure 1: Rules for Requirements Phrasing per Guideline

Name	Individual Compliance to Rules (5)-(39) from Table I
Target value	For each requirement r: Binary $[0, 1]$ where 1 means the quality rule is met. For a requirement set R: $[0 - 100]$ % $r \in R$ comply with the rule.
Quality factors	Unambiguous (all but rules (25), (26), & (33)), Appropriate (only rules (7) & (33)), Complete (only rules (12)-(14), (16)-(19), (28), (29), (31), & (34)-(39)), Singular (only rules (19), (24)-(27), & (36)), Verifiable (only rules (5)-(8), (10)-(20), (22)-(23), (27)-(28), (30)-(32), & (34)-(37)), Correct (only rules (9)-(12), (16), (21), (34), and (35)), & Conforming (all as guideline, explicitly mapped only rules (21) & (36))
Tools	Spreadsheet program (MS Excel)
Application	Check compliance to rules and detect bad smells.
Data items	Rule evaluation result $GR_j(r)$ for each requirement in the examined set $r \in R$ and each guideline rule $GR_j \mid j \in [5, \ldots, 39]$ as in Table I; $\#r_t$
Computation	$\%GR_j(R) = \frac{\#r_{GR_j}}{\#r_t} * 100, \ \ \#r_{GR_j} = \sum_{i=1}^{\#r_t} GR_j(r_i), \quad GR_j(r) = \begin{cases} 1, & \text{if the respective rule is satisfied,} \\ 0, & \text{else} \end{cases}$ Rules (13) & (14) can be combined to *"clearness of reference point"* [9] (German "Bezugspunkteindeutigkeit" (BPE)) and Rules (8) & (12) are part of *"clearness of process word"* [9] (German "Prozessworteindeutigkeit" (PE))
Interpretation	High numbers indicate many occurrences of the respective bad smell.
Considerations	The same calculations apply for general review results towards the specific quality factors. Too strict application of rules is criticized by some authors. In particular rules (5)+(8) [13], (22) [14], (24) [15, 16], (28) [13–16], and (34) [17].
Example	Let R consist of these two requirements from EagleEye [18]: (1) *"The AOCS subsystem shall account for redundancy of some hardware component to avoid critical and/or catastrophic consequences for the mission."* (2) *"The AOCS subsystem shall account for the following sensors: Star tracker, Three-axis gyros, Sun sensors, Magnetometers, GPS."* Evaluating $\%GR_j(R)$ for rule (25) "separate rationale from sentence": $GR_{25}(r_1) = 0, GR_{25}(r_2) = 1$ and $\%GR_{25}(R) = 50\%$
References	[1, 2, 9, 14, 19–24]

Table II: Binary Metrics for rules (5)-(39)

Name	Number of Sentences, Words, Process Verbs, or Punctuations
Target value	Natural number $\in \mathbb{N}_0\{0, 1, 2, \ldots\}$; critical values to meet the quality: - for sentences $\#s$ /process verbs $\#pv$: [1], - for words $\#w$: good $[5, \ldots, 15]$, medium $[16, \ldots, 20]$, - for punctuations $\#pt$: $< 209/1000$ words
Quality factors	Unambiguity, Comprehensibility, Verifiability (only rule (3)), Singularity (only rules (1) and (3)), and Conforming (as guidelines)
Tools	Spreadsheet program (MS Excel)
Application	Can be applied to an individual requirement wording or a whole set. Check compliance of individual requirements with rules (1)-(4) from Table I; give impression of phrasing complexity; use as auxiliary metrics within readability metrics, as defined in Table IV-VI.
Data items	String(s) of requirement wording(s).
Computation	$\#s(r), \#w(r), \#pv(r), \#pt(r) = \lvert S, W, PV, PT \rvert$, where $S, W, PV, PT = \{s, w, pv, pt \mid s, w, pv, pt \in r\}$ are sets of sentences s, words w, process verbs pv, and punctuation marks pt of the requirement r. Punctuations are normalized to 1000 words: $\#pt_{/1000w}(r) = \frac{\#pt(r)}{\#w(r)} * 1000$ For sets: $\#s(R), \#w(R), \#pv(R), \#pt(R) = \sum_{i=1}^{\#r_t} \#s(r_i), \#w(r_i), \#pv(r_i), \#pt(r_i)$ Thus, set average values can be calculated: $\varnothing s(R), \varnothing w(R), \varnothing pv(R), \varnothing pt(R) = \frac{\#s(R), \#w(R), \#pv(R), \#pt(R)}{\#r_t}$
Interpretation	Sentences should neither be too short to be complete nor too wordy, punctuations should be below average, and it should be exactly one sentence with one process verb per requirement - divergence from rules indicates a bad smell.
Considerations	Too strict application of rules is criticized by some authors. In particular rule (1) [13]. However, simpler and shorter sentences enhance readability. For readability measures see Table IV-VI.
Example	Let R consist of these two requirements from EagleEye [18]: (1) *"The AOCS subsystem shall account for redundancy of some hardware component to avoid critical and/or catastrophic consequences for the mission."* (2) *"The AOCS subsystem shall account for the following sensors: Star tracker, Three-axis gyros, Sun sensors, Magnetometers, GPS."* $\#s(r_1) = \#s(r_2) = 1,$ $\#w(r_1) = 20, \#w(r_2) = 17,$ $\#pv(r_1) = 2, \#pv(r_2) = 1,$ $\#pt(r_1) = 2, \#pt(r_2) = 6,$ $\#pt_{/1000w}(r_1) = 100, \#pt_{/1000w}(r_2) = 352.9$ and $\#pt_{/1000w}(R) = 216.2, \varnothing s(R) = 1, \varnothing w(R) = 18.5, \varnothing pv(R) = 1.5, \varnothing pt(R) = 4$
References	[1, 2, 19–21, 24–26]

Table III: Counting Metrics for rules (1)-(4)

<table>
<tr><td>Name</td><td colspan="3">Flesch Reading Ease Readability Score (FRE)</td></tr>
<tr><td rowspan="9">Target value</td><td colspan="3">Number rounded to Integer $\in [0, 1, \ldots, 100]$; critical values to meet the quality:</td></tr>
<tr><td>90–100</td><td>5th grade</td><td>Very easy to read. Easily understood by an average 11 year-old.</td></tr>
<tr><td>80–89</td><td>6th grade</td><td>Easy to read. Conversational English for consumers.</td></tr>
<tr><td>70–79</td><td>7th grade</td><td>Fairly easy to read.</td></tr>
<tr><td>60–69</td><td>8th-9th grade</td><td>Plain English. Easily understood by 13 to 15 year-olds.</td></tr>
<tr><td>50–59</td><td>10th-12th grade</td><td>Fairly difficult to read.</td></tr>
<tr><td>30–49</td><td>13th-16th grade (College)</td><td>Difficult to read.</td></tr>
<tr><td>10–29</td><td>College graduate</td><td>Very difficult to read.</td></tr>
<tr><td>0–9</td><td>Academic</td><td>Extremely difficult to read. Best understood by university graduates.</td></tr>
<tr><td>Quality factors</td><td colspan="3">Comprehensible</td></tr>
<tr><td>Tools</td><td colspan="3">Spreadsheet program (MS Excel), ReadabilityFormulas.com [27],
(Readable [28])</td></tr>
<tr><td>Application</td><td colspan="3">Determine the reading ease or complexity of a given text.</td></tr>
<tr><td>Data items</td><td colspan="3">Number of words $\#w(R)$, number of sentences $\#s(R)$, and number of syllables $\#sy(R)$ for the given set of requirements R. Although it is possible to calculate the formula for an individual requirement wording $r \in R$, it works best on samples of 100-300 words.</td></tr>
<tr><td>Computation</td><td colspan="3">$FRE(R) = 206.835 - 1.015 * \frac{\#w(R)}{\#s(R)} - 84.6 * \frac{\#sy(R)}{\#w(R)}$</td></tr>
<tr><td>Interpretation</td><td colspan="3">The higher the score, the lower the grade level respectively, the better, as this increases reading efficiency and reader persistence [29].</td></tr>
<tr><td>Considerations</td><td colspan="3">General appropriateness discussed in [29]. Original grade level to score mapping [30] is overlapping at interval boundaries and did not include separate academic level; all below 30 is college graduate. The weighting factors within the formula are based on language specific correlation statistics—here for English—and need to be adjusted for other languages. The formula targets “adult” reading and is not sensitive to differences in reading beginners texts $<$ 5th grade.</td></tr>
<tr><td>Example</td><td colspan="3">R = “The AOCS subsystem shall account for redundancy of some hardware component to avoid critical and/or catastrophic consequences for the mission.” from EagleEye [18]
$\#w(R) = 20,\ \#s(R) = 1,\ \#sy(R) = 40$
$FRE(R) = 206.835 - 1.015 * \frac{20}{1} - 84.6 * \frac{40}{20} \approx 17 \widehat{=}$ college graduate level</td></tr>
<tr><td>References</td><td colspan="3">[27, 29–33]</td></tr>
</table>

Table IV: Flesch Reading Ease Readability Score (FRE)

<table>
<tr><td>Name</td><td colspan="3">Dale-Chall Readability Formula (DC)</td></tr>
<tr><td rowspan="8">Target value</td><td colspan="3">Number; critical values to meet the quality:</td></tr>
<tr><td>≤ 4.9</td><td>4th grade & below</td><td>Very easy to read.</td></tr>
<tr><td>5.0–5.9</td><td>5th-6th grade</td><td>Easy to read.</td></tr>
<tr><td>6.0–6.9</td><td>7th-8th grade</td><td>Fairly easy to read.</td></tr>
<tr><td>7.0–7.9</td><td>9th-10th grade</td><td>Plain English.</td></tr>
<tr><td>8.0–8.9</td><td>11th-12th grade</td><td>Fairly difficult to read.</td></tr>
<tr><td>9.0–9.9</td><td>13th-15th grade (College)</td><td>Difficult to read.</td></tr>
<tr><td>≥ 10</td><td>College graduate</td><td>Very difficult to read.</td></tr>
<tr><td>Quality factors</td><td colspan="3">Comprehensible</td></tr>
<tr><td>Tools</td><td colspan="3">Spreadsheet program (MS Excel), ReadabilityFormulas.com [27], (Readable [28])</td></tr>
<tr><td>Application</td><td colspan="3">Determine the reading ease or complexity of a given text.</td></tr>
<tr><td>Data items</td><td colspan="3">Number of words $\#w(R)$, number of sentences $\#s(R)$, and number of “difficult” words $\#w_d(R)$ for the given set of requirements R. A word w is difficult if $w \notin L_{DC}$, where L_{DC} is a list of commonly known words according to [34]. Although it is possible to calculate the formula for an individual requirement wording $r \in R$, it works best on samples of 100-300 words.</td></tr>
<tr><td>Computation</td><td colspan="3">$DC_{raw}(R) = 15.79 * \frac{\#w_d(R)}{\#w(R)} + 0.0496 * \frac{\#w(R)}{\#s(R)}$
$DC(R) = \begin{cases} DC_{raw}(R) + 3.6365, & \text{if } \frac{\#w_d(R)}{\#w(R)} * 100 > 5, \\ DC_{raw}(R), & \text{else} \end{cases}$</td></tr>
<tr><td>Interpretation</td><td colspan="3">The lower the score, the lower the grade level respectively, the better, as this increases reading efficiency and reader persistence [29].</td></tr>
<tr><td>Considerations</td><td colspan="3">General appropriateness discussed in [29]. The weighting factors within the formula are based on language specific correlation statistics—here for English—and need to be adjusted for other languages. The formula targets “adult” reading and is not sensitive to differences in reading beginners texts $<$ 5th grade.</td></tr>
<tr><td>Example</td><td colspan="3">R = “The AOCS subsystem shall account for redundancy of some hardware component to avoid critical and/or catastrophic consequences for the mission.” from EagleEye [18]
$\#w(R) = 20,\ \#s(R) = 1,\ \#w_d(R) = 8,\ \frac{\#w_d(R)}{\#w(R)} * 100 = 40 > 5$
$DC(R) = 0.1579 * \frac{8}{20} + 0.0496 * \frac{20}{1} + 3.6365 = 10.9 \widehat{=}$ college graduate level</td></tr>
<tr><td>References</td><td colspan="3">[27, 29, 32, 34]</td></tr>
</table>

Table V: Dale-Chall Readability Formula (DC)

Name	Grade Level Reading Metrics a) Flesch-Kincaid Grade Level (FK) [35] b) Gunning Fog Index (GFI) [36] c) SMOG Index [37] d) Coleman-Liau Index (CLI) [38] e) Automated Readability Index (ARI) [35] f) Linsear Write (LW) [27, 39] g) Fry Readability Graph [40] h) Raygor Estimate Graph [41]
Target value	Number > 0 estimating years of education necessary to understand the text; critical values to meet the quality: < 5 Reading beginners. Formulas not optimized for these levels. 5 Very easy to read. Easily understood by an average 11 year-old. 6 Easy to read. Conversational English for consumers. 7 Fairly easy to read. 8-9 Plain English. Easily understood by 13 to 15 year-olds. 10-12 Fairly difficult to read. 13-16 Difficult to read. College level. > 16 Very difficult to read. College or university graduates.
Quality factors	Comprehensible
Tools	Spreadsheet program (MS Excel), ReadabilityFormulas.com [27], (Readable [28])
Application	Determine the reading ease or complexity of a given text.
Data items	Number of words $\#w(R)$, number of sentences $\#s(R)$, number of syllables $\#sy(R)$, number of letters $\#l(R)$, number of charters (letters and numbers) $\#c(R)$, and number of polysyllabic words $\#w_{\#sy(w)\geq x}(R)$ with $x = 3$ for the given set of requirements R. For $\#w_{\#sy(w)\geq x}(R)$, proper names, combinations of easy words, and verbs enlonged by suffixes as -ed, -es, or -ing are ignored. Although it is possible to calculate the formulas for an individual requirement wording $r \in R$, they work best on samples of 100-300 words.
Computation	a) $FK(R) = 0.39 * \frac{\#w(R)}{\#s(R)} + 11.8 * \frac{\#sy(R)}{\#w(R)} - 15.59$ b) $GFI(R) = 0.4 * (\frac{\#w(R)}{\#s(R)} + 100 * \frac{\#w_{\#sy(w)\geq 3}(R)}{\#w(R)})$ c) $SMOG(R) = 1.043 * \sqrt{30 * \frac{\#w_{\#sy(w)\geq 3}(R)}{\#s(R)}} + 3.1291$ d) $CLI(R) = 5.88 * \frac{\#l(R)}{\#w(R)} - 29.6 * \frac{\#s(R)}{\#w(R)} - 15.8$ e) $ARI(R) = 4.71 * \frac{\#c(R)}{\#w(R)} + 0.5 * \frac{\#w(R)}{\#s(R)} - 21.43$ f) $LW_{raw}(R) = \frac{\#w_{\#sy(w)\leq 2}(R) + 3 * \#w_{\#sy(w)\geq 3}(R)}{\#s(R)}$, $LW(R) = \begin{cases} LW_{raw}(R)/2, & \text{if } LW_{raw}(R) > 20, \\ (LW_{raw}(R) - 2)/2, & \text{else} \end{cases}$ g) $Fry(R) = lookup_{FryGraph}(\frac{\#s(R)}{\#w(R)} * 100, \frac{\#sy(R)}{\#w(R)} * 100)$ h) $Raygor(R) = lookup_{RaygorGraph}(\frac{\#s(R)}{\#w(R)} * 100, \frac{\#w_{\#c\geq 6}(R)}{\#w(R)} * 100)$
Interpretation	The lower the grade level, the better, as this increases reading efficiency and reader persistence [29].
Considerations	General appropriateness discussed in [29, 32, 42]. Weighting factors within the formulas optimized for English. Other languages need adjustment. The formulas target "adult" reading and are not sensitive to differences in reading beginners texts $<$ 5th grade.
Example	*R = "The AOCS subsystem shall account for redundancy of some hardware component to avoid critical and/or catastrophic consequences for the mission." from EagleEye [18]* $\#w(R) = 20$, $\#s(R) = 1$, $\#sy(R) = 40$, $\#l(R) = 121 = \#c(R)$, $\#w_{\#sy(w)\geq 3}(R) = 6$, $\#w_{\#sy(w)\leq 2}(R) = 14$, $\#w_{\#c\geq 6}(R) = 10$ a) $FK(R) = 0.39 * \frac{20}{1} + 11.8 * \frac{40}{20} - 15.59 = 15.81 \widehat{=}$ college level b) $GFI(R) = 0.4 * (\frac{20}{1} + 100 * \frac{6}{20}) = 20 \widehat{=}$ college graduate level c) $SMOG(R) = 1.043 * \sqrt{30 * \frac{6}{1}} + 3.1291 \approx 17 \widehat{=}$ college graduate level d) $CLI(R) = 5.88 * \frac{121}{20} - 29.6 \frac{1}{20} - 15.8 = 18.29 \widehat{=}$ college graduate level e) $ARI(R) = 4.71 * \frac{121}{20} + 0.5 * \frac{20}{1} - 21.43 \approx 17 \widehat{=}$ college graduate level f) $LW(R) = \frac{14 + 3 * 6}{1}/2 = 15 \widehat{=}$ college level g) $Fry(R) = lookup_{FryGraph}(\frac{1}{20} * 100 = 5, \frac{40}{20} * 100 = 200) \widehat{=}$ invalid h) $Raygor(R) = lookup_{RaygorGraph}(\frac{1}{20} * 100 = 5, \frac{10}{20} * 100 = 50) \widehat{=}$ invalid
References	[24, 27, 29, 32, 33, 35–38, 40–43]

Table VI: Grade Level Readability Formulas

Name	Estimated Reading Time
Target value	Decimal number referring to number of minutes - can be transformed to any time format. There is not absolute critical value, the measure is used relative to compare different results.
Quality factors	Efficiency
Tools	Spreadsheet program (MS Excel), (Readable [28])
Application	Measure how long it takes to read the specification.
Data items	String(s) of requirement wording(s) $r \in R$ and their number of words $\#w(R)$.
Computation	$RT(R) = \frac{\#w(R)}{200} \qquad \varnothing RT(R) = \frac{RT(R)}{\#r_t(R)}$
Interpretation	Faster reading is better. However, absolute reading time depends on length of specification. To compare different specifications the average per requirement should be compared.
Considerations	The formula directly depends on number of words $\#w$. Yet, time is a measure more intelligible in terms of efficiency. Practical reading time depends on reading ease and its fit with the readers capacities. For readability measures see Table IV-VI. However, average reading time gives impression of time effort needed to process the text in general. Time can also be measured experimentally with test subjects, not only for reading, but also for writing. In general, time is a common efficiency measure [44].
Example	Let R consist of these two requirements from EagleEye [18]: (1) *"The AOCS subsystem shall account for redundancy of some hardware component to avoid critical and/or catastrophic consequences for the mission."* (2) *"The AOCS subsystem shall account for the following sensors: Star tracker, Three-axis gyros, Sun sensors, Magnetometers, GPS."* $\#w(r_1) = 20, \#w(r_2) = 17,$ $RT(r_1) = 6sec, RT(r_2) = 5sec, \varnothing RT(R) = 5.5sec$
References	[44–46]

Table VII: Estimated Reading Time

Name	F-Score
Target value	Percentage of formality within 0 - 100% critical values are unknown due to lack of comparison values.
Quality factors	Formality
Tools	Spreadsheet program (MS Excel), custom Python tool [47]
Application	Measure *deep formality* of the text (level of context needed to understand).
Data items	String(s) of requirement wording(s) $r \in R$ and their percentage of words belonging to a specific category or part of speech (POS) — noun (NN), verb (VB), article (AT), adjective (JJ), preposition (IN), pronoun (PN), adverb (RB), and interjection (UH) $\%w_i(R) = \frac{\#w_i(R)}{\#w(R)} * 100$ with $i \in NN, VB, AT, JJ, IN, PN, RB, UH$.
Computation	$F{-}Score(R) = 50 + \frac{\%w_{NN}(R) + \%w_{JJ}(R) + \%w_{IN}(R) + \%w_{AT}(R)}{2} - \frac{\%w_{PN}(R) + \%w_{VB}(R) + \%w_{RB}(R) + \%w_{UH}(R)}{2}$
Interpretation	Higher numbers correspond to less context and thus are better. Yet, reference values are missing, in particular for requirements. Results in related work for different genres range from -55-70% [48, 49]. Thus, values above 40% are expected, but in general the comparison is the goal not the absolute numbers.
Considerations	Discussion on performance in [48]. Works better on larger samples.
Example	*R = "The AOCS subsystem shall account for redundancy of some hardware component to avoid critical and/or catastrophic consequences for the mission." from EagleEye [18]* $\#w(R) = 20, \%w_{NN}(R) = 35, \%w_{JJ}(R) = 10, \%w_{IN}(R) = 20, \%w_{AT}(R) = 10, \%w_{PN}(R) = 5, \%w_{VB}(R) = 15, \%w_{RB}(R) = 0, \%w_{UH}(R) = 0,$ $F - Score(R) = 50 + \frac{(35 + 10 + 20 + 10) - (5 + 15 + 0 + 0)}{2} = 77.5$
References	[48–51]

Table VIII: F-Score Formality Measure

Figure 2: Metrics for Quality Attributes of Requirements Phrasings

Name	Requirement Phrasings.
Metrics	Guideline Based Metrics (Table II & III), Readability Scores (Table IV-VI), Reading-, Writing-, Review-Time (Table VII), F-Score (Table VIII), and Subjective Readability, Learnability, & Quality (questionnaire).
Definition	Phrasings of requirements in different template notations.
Source	Rephrased from original documents [52–55].
Collector	Researchers and research assistants, in some cases test subjects.
Timing	Before or during experiments.
Procedures	Manual rephrasing through expert or test subject.
Storage	Spreadsheet.
Representation	Textual.
Sample	Select requirement documents as representative for the targeted domain(s) and abstraction level(s). Include all requirements of the document, if possible.
Verification	Cross-checking through experts or template compliance checking tool.
Alternatives	-
Integrity	Phrasings from user experiments are not to be changed. Expert phrasings as input to experiments can be changed after cross checking quality assessment and discussion.

Table IX: Data Item Definition for Requirement Phrasings

Name	Requirement Quality Assessment.
Metrics	Guideline Based Metrics (Table II & III), Readability Scores (Table IV-VI), Reading-Time (Table VII), F-Score (Table VIII).
Definition	Binary quality assessment or key data on text characteristics of requirements phrasings necessary to calculate metrics.
Source	Table IX.
Collector	Researchers (and research assistants), in some cases test subjects.
Timing	Before the measurement of expressiveness.
Procedures	Manual assessment and where possible automated by spreadsheet formula or light weight natural language processing.
Storage	Spreadsheet.
Representation	Matrix requirement:characteristic, binary characteristics (Table II) boolean as [1,0], others (Table III) numeric count.
Sample	Phrasings selected for Table IX. Characteristics as specified in Table II & III.
Verification	Sample inspection though and discussion with other researchers/experts.
Alternatives	Fully automated through natural language processing.
Integrity	Phrasings from user experiments are not to be changed. Expert phrasings as input to experiments can be changed after cross checking quality assessment and discussion.

Table X: Data Item Definition for Requirement Quality Assessment

Wolf and Strößner's [9] unambiguity metric can be calculated as a secondary metric from our results:

$$Unambiguity = \frac{u * PE + v * BPE + w * BE}{\#r_t} * 100, \text{ where}$$

$\#r_t$ is the total number of requirements in the examined set,

PE is the unambiguity of the process words (Ger. "Prozesswortеindeutigkeit") that is the count of all requirements phrased in active voice and using a full verb—a precise verb that is no nominalization and no light verb construction,

BPE is the unambiguity of the reference points (Ger. "Bezugspunkteindeutigkeit") that is the count of all requirements that contain no comparison or where the comparison is clear,

BE is the term unambiguity (Ger. "Begriffseindeutigkeit") that is the count of all requirements where all terms are clear and defined, e.g., in a glossary, and

u, v, w are factors to weight these for the project context.

As term definitions are irrelevant to our experimentation goals, we assume $w = 0$. PE and BPE can be calculated from individual metric evaluations per requirement. Further, as we have no context that provides reason to weight both values, we assume $u = v = 0.5$.

IV. Data Items

The different metrics, as introduced above, are applied in different experiments to requirements phrased following different template systems. This data item is summarized in Table IX following the data item template from IEEE 1061 [12]. Table X describes in the same way the individual quality ratings of requirements as a data item.

Acknowledgment

omitted for double-blind review

References

[1] Requirements Working Group. *Guide for Writing Requirements*. Tech. rep. INCOSE-TP-2010-006-03. Version 3. International Council on Systems Engineering (INCOSE), July 19, 2019.

[2] Chris Rupp and Andreas Günther. "Das SOPHIST-REgelwerk - Psychotherapie für Anforderungen". German. In: Chris Rupp and SOPHIST GmbH. *Requirements-Engineering und -Management – Aus der Praxis von klassisch bis agil*. 6th ed. Carl Hanser Verlag München, 2014, pp. 123–164. ISBN: 978-3-446-43893-4.

[3] ECSS Secretariat and ESA-ESTEC Requirements & Standards Division. *Space engineering - Technical requirements specification*. ECSS-E-ST-10-06C (ECSS). Mar. 6, 2009.

[4] ECSS Secretariat and ESA-ESTEC Requirements & Standards Division. *ECSS - Draft rules and template for ECSS Standards*. ECSS-D-00-01C (ECSS). May 20, 2014.

[5] *ISO/IEC/IEEE 29148: Systems and software engineering – Life cycle processes – Requirements engineering*. ISO/IEC/IEEE 29148:2018(E) (ISO/IEC/IEEE). Nov. 2018.

[6] Michael Alexander et al. *NASA SYSTEMS ENGINEERING HANDBOOK*. Tech. rep. NASA SP-2016-6105 Rev2. NASA, 2016. URL: https://www.nasa.gov/connect/ebooks/nasa-systems-engineering-handbook (visited on 09/10/2021).

[7] Chris Rupp and Rainer Joppich. "Anforderungsschablonen - der MASTER-Plan für gute Anforderungen". German. In: Chris Rupp and SOPHIST GmbH. *Requirements-Engineering und -Management – Aus der Praxis von klassisch bis agil*. 6th ed. Carl Hanser Verlag München, 2014, pp. 215–246. ISBN: 978-3-446-43893-4.

[8] Chris Rupp and Stefan Queins. "Von der Idee zur Spezifikation". German. In: Chris Rupp and SOPHIST GmbH. *Requirements-Engineering und -Management – Aus der Praxis von klassisch bis agil*. 6th ed. Carl Hanser Verlag München, 2014, pp. 33–50. ISBN: 978-3-446-43893-4.

[9] Ellen Wolf and Matthias Strößner. "Qualitätsmetriken". German. In: Chris Rupp and SOPHIST GmbH. *Requirements-Engineering und -Management. Professionelle, iterative Anforderungsanalyse für die Praxis*. 5th ed. Carl Hanser Verlag GmbH und Co. KG, 2009, pp. 313–339. ISBN: 978-3-44641-841-7.

[10] Die SOPHISTen. *MASTER - Schablonen für alle Fälle*. German. Ed. by Roland Kluge. 2016. URL: https://www.sophist.de/fileadmin/user_upload/Bilder_zu_Seiten/Publikationen/Wissen_for_free/MASTeR_Broschuere_3-Auflage_interaktiv.pdf (visited on 09/24/2019).

[11] Chris Rupp. *Requirements Templates - The Blueprint of your Requirement*. Requirements-Engineering und -Management - 6. Auflage - Webinhalte zu Kapitel 10. 2014. URL: https://www.sophist.de/re6/webinhalte-buchteil-iii/ (visited on 11/04/2016).

[12] *IEEE Standard for a Software Quality Metrics Methodology*. IEEE 1061-1998 (IEEE). Dec. 1998.

[13] Maxime Warnier and Anne Condamines. "A Case Study on Evaluating the Relevance of Some Rules for Writing Requirements through an Online Survey". In: *25th IEEE International Requirements Engineering Conference (RE'17)*. 2017, pp. 243–252. DOI: 10.1109/RE.2017.11.

[14] Henning Femmer et al. "Rapid Requirements Checks with Requirements Smells: Two Case Studies". In: *1st International Workshop on Rapid Continuous Software Engineering (RCoSE'14)*. 2014, pp. 10–19. DOI: 10.1145/2593812.2593817.

[15] Anne Condamines and Maxime Warnier. "Linguistic Analysis of Requirements of a Space Project and Their Conformity with the Recommendations Proposed by a Controlled Natural Language". In: *4th International Workshop Controlled Natural Language (CNL)*. Ed. by Brian Davis, Kaarel Kaljurand, and Tobias Kuhn. 2014, pp. 33–43. DOI: 10.1007/978-3-319-10223-8_4.

[16] Maxime Warnier. "How can corpus linguistics help improve requirements writing? Specifications of a space project as a case study". In: *23rd IEEE International Requirements Engineering Conference (RE'15)*. 2015, pp. 388–392. DOI: 10.1109/RE.2015.7320456.

[17] Katharina Winter, Henning Femmer, and Andreas Vogelsang. "How Do Quantifiers Affect the Quality of Requirements?" In: *26th International Working Conference on Requirements Engineering: Foundation for Software Quality (REFSQ'20)*. Ed. by Nazim Madhavji et al. 2020, pp. 3–18. DOI: 10.1007/978-3-030-44429-7_1.

[18] Francesco Pace. *EARTH OBSERVATION REFERENCE MISSION - SYSTEM SPECIFICATION*. Tech. rep. ATB-RAC-D5. ESA - ESTEC, 2009.

[19] Matthias Strößner and Thorsten Cziharz. "Qualitätsmetriken. drum messe, wer sich ewig bindet". German. In: Chris Rupp and SOPHIST GmbH. *Requirements-Engineering und -Management. Aus der Praxis von klassisch bis agil*. 6th ed. Carl Hanser Verlag München, 2014, pp. 301–316. ISBN: 978-3-446-43893-4.

[20] Alessio Ferrari et al. "Detecting requirements defects with NLP patterns: an industrial experience in the railway domain". In: *Empirical Software Engineering* 23.6 (Dec. 2018), pp. 3684–3733. ISSN: 1573-7616. DOI: 10.1007/s10664-018-9596-7.

[21] The Reuse Company. *RQA - Quality Studio*. 2019. URL: https://www.reusecompany.com/rqa-quality-studio (visited on 11/05/2019).

[22] Mohammed Javeed Ali. "Metrics for Requirements Engineering". MA thesis. Umeå University, 2006.

[23] Shahid Iqbal and M. Naeem Ahmed Khan. "Yet another Set of Requirement Metrics for Software Projects". In: *International Journal of Software Engineering and Its Applications* 6.1 (2012), pp. 19–28.

[24] Giuseppe Lami et al. "QuARS: Automated Natural Language Analysis of Requirements and Specifications". In: *INCOSE International Symposium* 15.1 (2005), pp. 344–353. DOI: 10.1002/j.2334-5837.2005.tb00674.x.

[25] Alessio Ferrari, Giorgio Oronzo Spagnolo, and Stefania Gnesi. "PURE: A Dataset of Public Requirements Documents". In: *25th IEEE International Requirements Engineering Conference (RE'17)*. 2017, pp. 502–505. DOI: 10.1109/RE.2017.29.

[26] Vivian Cook. "Standard Punctuation and the Punctuation of the Street". In: *Essential Topics in Applied Linguistics and Multilingualism: Studies in Honor of David Singleton*. Ed. by Mirosław Pawlak and Larissa Aronin. Springer International Publishing, 2014, pp. 267–290. DOI: 10.1007/978-3-319-01414-2_16. URL: http://www.viviancook.uk/Punctuation/PunctFigs.htm (visited on 10/26/2021).

[27] Brian Scott. *Readability Formulas. Free readability tools to check for Reading Levels, Reading Assessment, and Reading Grade Levels.* URL: https://readabilityformulas.com (visited on 10/28/2021).

[28] Dave Child. *Readable*. (formerly readable.io). URL: https://readable.com (visited on 10/28/2021).

[29] William H. DuBay. *The Principles of Readability*. Aug. 25, 2004. URL: https://eric.ed.gov/?id=ed490073.

[30] Rudolph F. Flesch. *The art of readable writing*. Harper Collins, New York, 1949. As cited in: William H. DuBay. *The Principles of Readability*. Aug. 25, 2004. URL: https://eric.ed.gov/?id=ed490073.

[31] Rudolph Flesch. "A new readability yardstick." In: *Journal of Applied Psychology* 32.3 (1948), pp. 221–233. DOI: 10.1037/h0057532.

[32] George R. Klare. "Assessing Readability". In: *Reading Research Quarterly* 10.1 (1974), pp. 62–102. DOI: 10.2307/747086.

[33] William M. Wilson, Linda H. Rosenberg, and Lawrence E. Hyatt. "Automated Analysis of Requirement Specifications". In: *19th IEEE International Conference on Software Engineering (ICSE'97)*. May 1997, pp. 161–171. DOI: 10.1145/253228.253258.

[34] Jeanne Sternlicht Chall and Edgar Dale. *Readability revisited: The new Dale-Chall readability formula*. Brookline Books, 1995. As cited in: Brian Scott. *Readability Formulas. Free readability tools to check for Reading Levels, Reading Assessment, and Reading Grade Levels.* URL: https://readabilityformulas.com (visited on 10/28/2021).

[35] J. Peter Kincaid et al. *Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel*. Tech. rep. Research Branch Report 8-75. U.S. Naval Technical Training Command, Naval Air Station Memphis - Millington, TN, Feb. 1975.

[36] Robert Gunning. *The technique of clear writing*. McGraw-Hill, New York, 1968. As cited in: Judith Bogert. "In Defense of the Fog Index". In: *The Bulletin (of the Association for Business Communication)* 48.2 (June 1985), pp. 9–12. DOI: 10.1177/108056998504800203.

[37] G. Harry McLaughlin. "SMOG Grading - a New Readability Formula". In: *Journal of Reading* 12.8 (1969), pp. 639–646. URL: http://www.jstor.org/stable/40011226.

[38] Meri Coleman and T. L. Liau. “A computer readability formula designed for machine scoring.” In: *Journal of Applied Psychology* 60.2 (1975), pp. 283–284. DOI: 10.1037/h0076540.
[39] John O’Hayre. *Gobbledygook Has Gotta Go*. U.S. Department of the Interior, Bureau of Land Management, 1966.
[40] Edward Fry. “A Readability Formula That Saves Time”. In: *Journal of Reading* 11.7 (1968), pp. 513–578. ISSN: 00224103. URL: http://www.jstor.org/stable/40013635.
[41] Alton L. Raygor. “The Raygor readability estimate: A quick and easy way to determine difficulty”. In: National Reading Conference Clemson, SC, 1977, pp. 259–263. As cited in: R. Scott Baldwin and Rhonda K. Kaufman. “A Concurrent Validity Study of the Raygor Readability Estimate”. In: *Journal of Reading* 23.2 (1979), pp. 148–153.
[42] Judith Bogert. “In Defense of the Fog Index”. In: *The Bulletin (of the Association for Business Communication)* 48.2 (June 1985), pp. 9–12. DOI: 10.1177/108056998504800203.
[43] R. Scott Baldwin and Rhonda K. Kaufman. “A Concurrent Validity Study of the Raygor Readability Estimate”. In: *Journal of Reading* 23.2 (1979), pp. 148–153.
[44] Kasper Hornbæk. “Current practice in measuring usability: Challenges to usability studies and research”. In: *International Journal of Human-Computer Studies* 64.2 (Feb. 2006), pp. 79–102. DOI: 10.1016/j.ijhcs.2005.06.002.
[45] Marc Brysbaert. “How many words do we read per minute? A review and meta-analysis of reading rate”. In: *Journal of Memory and Language* 109 (2019), p. 104047. ISSN: 0749-596X. DOI: 10.1016/j.jml.2019.104047.
[46] Cris Trauntner. *How to Calculate Reading Time*. Infusionmedia. Sept. 24, 2020. URL: https://infusion.media/content-marketing/how-to-calculate-reading-time/ (visited on 11/10/2021).
[47] Anonymous. *Evaluation of templates for requirements documentation*. Mar. 16, 2023. DOI: 10.5281/zenodo.6321277.
[48] Haiying Li, Zhiqiang Cai, and Arthur C. Graesser. “Comparing Two Measures for Formality”. In: *Twenty-Sixth International Florida Artificial Intelligence Research Society Conference*. 2013, pp. 220–225.
[49] Francis Heylighen and Jean-Marc Dewaele. “Variation in the contextuality of language: an empirical measure”. In: *Foundations of Science* 7.3 (2002), pp. 293–340. DOI: 10.1023/A:1019661126744.
[50] Francis Heylighen and Jean-Marc Dewaele. *Formality of language: definition, measurement and behavioral determinants*. Internal Report. Center ”Leo Apostel”, Free University of Brussels, 1999.
[51] Daniel Eriksson. “Using the F-measure to test formality in sports reporting. A comparison of the language used in soccer and horse polo articles in two British newspapers”. MA thesis. Karlstad University, Department of Language, Literature and Intercultural Studies, 2013.
[52] *Certification Specifications for Engines*. CS-E, Amendment 1, Annex to ED Decision 2007/015/R (European Aviation Safety Agency (EASA)). Dec. 10, 2007. URL: https://www.easa.europa.eu (visited on 10/14/2021).
[53] ECSS Secretariat and ESA-ESTEC Requirements & Standards Division. *Space engineering - Satellite attitude and orbit control system (AOCS) requirements*. ECSS-E-ST-60-30C (ECSS). Aug. 20, 2013.
[54] FLEX Team. *FLEX Space Segment Requirements Document (SSRD)*. Tech. rep. FLX-RS-ESA-SYS-0042. Version 1.1. ESA - ESTEC, Apr. 24, 2017.
[55] Shurouq Abusalah et al. *NBDiff 1 documentation: Software Requirements Specification*. 2014. URL: https://nbdiff-docs.readthedocs.io/en/latest/SRS.html (visited on 11/23/2021).