Measuring User Rated Language Quality: Development and Validation of the User Interface Language Quality Survey (LQS)

Javier A. Bargas-Avila[a,*], Florian Brühlmann[a]

[a] *Google / YouTube User Experience Research, Brandschenkestrasse 110, 8002 Zurich, Switzerland*

Abstract

Written text plays a special role in user interfaces. Key information in interaction elements and content are mostly conveyed through text. The global context, where software has to run in multiple geographical and cultural regions, requires software developers to translate their interfaces into many different languages. This translation process is prone to errors – therefore the question of how language quality can be measured is important. This article presents the development of a questionnaire to measure user interface language quality (LQS). After a first validation of the instrument with 843 participants, a final set of 10 items remained, which was tested again (N = 690). The survey showed a high internal consistency (Cronbach's α) of .82, acceptable discriminatory power coefficients (.34 – .47), as well as a moderate average homogeneity of .36. The LQS also showed moderate correlation to UMUX, an established usability metric (convergent validity), and it successfully distinguished high and low language quality (discriminative

*Corresponding author
Email addresses: javier.bargas@me.com (Javier A. Bargas-Avila), florian.bruehlmann@gmail.com (Florian Brühlmann)

validity). The application to three different products (YouTube, Google Analytics, Google AdWords) revealed similar key statistics, providing evidence that this survey is product-independent. Meanwhile, the survey has been translated and applied to more than 60 languages.
Keywords: User Interface, Language, Text, Translation, Internationalization, Localization, l10n, i18n

---

1. Introduction

Key information in interaction elements and content within user interfaces are mostly conveyed through text. Graphical user interfaces have evolved substantially when compared to text-based user interfaces, but they still rely heavily on language to communicate with users. Therefore language plays a crucial role in Human-Computer Interaction. Single words can make the difference between failure or success.

The importance of language within a user interface (UI) becomes clear when text elements are removed. Figure 1 shows three screenshots of the video-sharing site YouTube. The first (a) shows the original, the second (b) shows the website, but with all text elements removed, while on the third (c) all graphic elements are deleted. The illustration shows how the textless version is stripped of the most useful information: It is almost impossible to predict and choose which video to watch and navigation becomes impossible.

Text used in interfaces is highly dependent on cultural and regional aspects. For example, instructional text such as a tutorial could be worded informally for the US, but such an informal wording might be very inappropriate in other cultures. Hence it is important to consider not only mere

correctness of translation of text but also style and tone aspects in the specific cultural context. Beside translation of text, interface elements such as icons and pictures should also be considered in the process of localization. Worldwide, there are about 200 languages that are spoken by at least 3 million people (Lewis et al., 2013). Companies with worldwide reach need to localize their products to make sure they can be used by everyone. For instance, Google search currently supports more than 140, Facebook more than 60, and YouTube more than 60 languages.

Websites and user interfaces are generally developed in one source language and translated afterwards by professional linguists. The process of translation is prone to errors and might introduce a number of problems that are not present in the source user interface. For example, the word auto can be translated to French as *automatique* (automatic) or *automobile* (car), which obviously has a completely different meaning. Another problem arises from words that behave as a verb when placed in a button or as a noun if part of a label (Leiva and Alabau, 2014). For example, the word *access* can stand for "you have access" (as a label) or "you can request access" (as a button). This *word sense disambiguation problem* (Muntés Mulero et al., 2012) arises often in UI translations. Further, possible pitfalls are gender, prepositions without context (Muntés Mulero et al., 2012) or other characteristics of the source text that might influence the translation process (Dilts, 2001). Such mistranslations might not only negatively affect trustworthiness and brand perception, but also the acceptance of the website and its perceived usefulness (Sun, 2001).

As companies scale their products to multiple languages, the need for

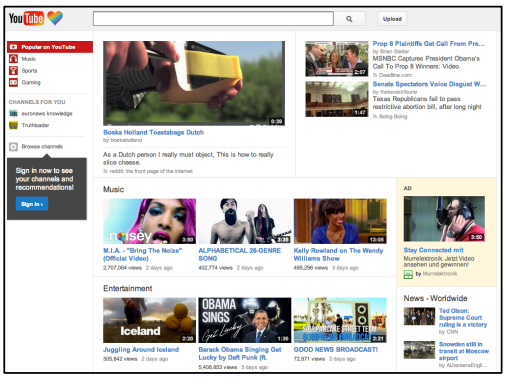
quality metrics increases: How can product managers learn more about the quality of a translation in an interface when they might not even speak the language themselves? In this paper, a method is presented that delivers metrics about language quality by asking users to rate the language of the user interfaces in a survey.

2. Theoretical Background

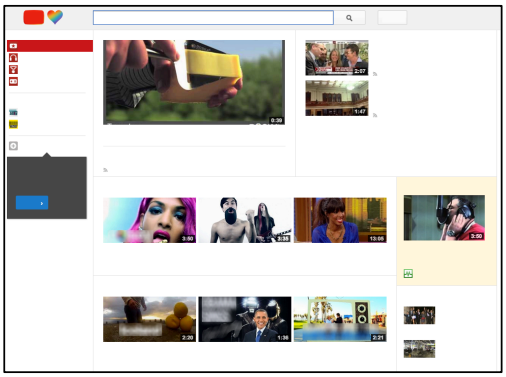
Schriver (1989) distinguishes three different classes of text quality evaluation: 1) text-focused, 2) expert-judgement-focused, and 3) reader-focused. These three classes express different levels on how explicit the feedback from the target audience is: "...text-focused methods (...) never use direct reader response; experts – through their experience – provide surrogate reader feedback; and reader-focused methods make explicit use of audience response." (Schriver, 1989, p. 241).

2.1. Text-focused evaluation

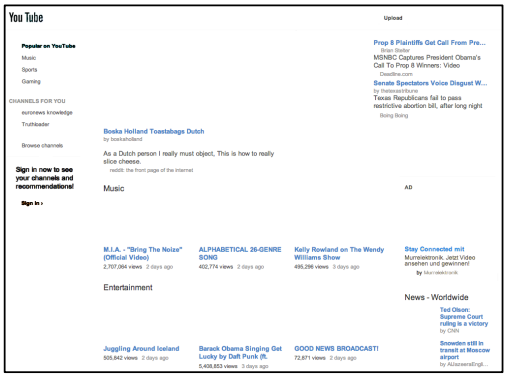
Text-focused methods operate by having a person or a computer examine a text and assess text quality by applying rules and guidelines that define what good text quality is. These methods include readability formulae (e.g., Fry, 1968; Kincaid et al., 1975) and user models (e.g., Blackmon et al., 2005; Chi et al., 2001) which can be applied by software that would allow automation of certain aspects of evaluation. Such automized analysis is inexpensive and can spot certain obvious classes of error such as misspellings or provide general statistics about number of complex or passive sentences that could reduce readability. But in general, these provide little information about the



(a)



(b)



(c)

Figure 1: Example of how UIs look when text or graphics are removed.

overall performance of the text (whole-text level) or whether the text meets the needs of readers.

2.2. Expert-judgement-focused evaluation

Expert reviews involve a systematic screening of the text corpus by professional linguists. The major advantage of this method is that in-depth valuable feedback, which is based on expert knowledge, is produced. A drawback of this method can arise if evaluators are too close to the text or product that is examined, therefore making it harder to mentally take the users perspective when evaluating the language (Schriver, 1989). Also, this method is quite expensive to scale for products that are translated into many different languages.

2.3. Reader-focused evaluation

Schriver (1989) distinguishes two classes of reader feedback methods: (1) Concurrent tests that evaluate the behaviors of readers in real-time, and (2) retrospective tests that are usually applied shortly after the reader has finished reading the text or after a certain time period. Concurrent methods include performance testing and thinking-aloud methods, while retrospective methods involve comprehension tests and surveys. Retrospective user testing is useful for revising existing text (Schriver, 1989).

Reader-focused methods have the advantage of giving information on global aspects of text quality and information about how the audience may respond to the text (Schriver, 1989). While retrospective methods such as surveys have disadvantages over concurrent methods (e.g., thinking-aloud or performance testing) because they rely on the use of memory, a survey during

or after the interaction with a software might be a relatively reliable method to measure text quality. An empirical comparison of expert-focused and reader-focused methods of text evaluation showed that mutual agreement on problems in a text among experts is usually relatively low and contributed to a large set of false-alarms – problems that the readers did not report (Lentz and de Jong, 1997). This study also showed that experts experience difficulties with predicting the problems that readers reported. The feedback of users is thus invaluable for judging the quality perception of text.

Schriver (1989) argues that expert-judgement-focused evaluation should be used in combination with reader-focused evaluation methods to ensure the text comprehension of the target audience.

2.4. Background: How to focus on quality assurance resources?

In 2012, the YouTube internationalization team was in the following situation: Anecdotal evidence suggested that some language versions of YouTube might benefit from improvement efforts. Past projects had shown that expert evaluations yielded good results and led to significant improvements of text quality. The problem with these evaluations was that they were time- and resource-consuming to conduct and analyze. The team did not have enough resources to conduct these reviews for all 60 languages and needed a reliable method to understand the state of each version.

User interface text is one among many aspects, such as date formats, color or icons and symbols, that need to be considered in the localization of a product. While there are guidelines for internationalization such as those proposed by Del Galdo (1990), there are, to the authors knowledge, no validated scales available to specifically evaluate UI text quality. Nielsen

(1990) argues that a localized interface should be regarded as a *new* interface and therefore tested and analyzed accordingly. While task-based user testing of localized interfaces is important, users might not provide feedback about the language quality that goes beyond text errors encountered during a task. Also, usability testing with users for more than 60 language versions of an interface is very expensive and time consuming.

Based on this situation it was decided to apply a reader-focused method, and have YouTube users provide feedback on language quality through a survey. These data would then be used to determine which languages should be improved by expert evaluation efforts.

2.5. Six-subgroup quality scale

To the authors knowledge there is only one published scale that measures perceived text quality. The Six-Subgroup Quality Scale (SSQS) supports reviewers during the evaluation of an essay (Ransdell and Levy, 1996). It consists of six dimensions: 1) Words: Choice and arrangement (readability), 2) Technical quality: Mechanics (tenses, grammar, spelling), 3) Content of essay (engagement, egocentrism), 4) Purpose/audience/tone (clear purpose, language and tone), 5) Organization and development (elaboration, completeness, paragraphing), and 6) Style (sentence structure, creativity).

While these quality criteria make sense for the evaluation of a multi-paragraph essay, not all of these aspects are relevant for user interface text. Many user interface text segments consist only of one word or a sentence. Applying, for instance, the categories "Content of essay", "Organization and development" or "style" on user interface strings would yield little useful data.

Due to this situation, it was decided to develop and validate a survey to measure user interface language quality. The Language Quality Survey (short: LQS) aims to facilitate feedback for researchers and practitioners about the text quality of user interfaces and enable focused quality improvement efforts on problematic languages.

Note that this publication reports the development and validation of this survey. It does not report detailed results and findings regarding YouTube's language quality.

3. Development and first validation

3.1. Development of the LQS

3.1.1. Item-generation for the first version

In a first step, a group of professional linguists came together in a brainstorming session and discussed the core criteria of language quality. These linguists were experts in their field and involved in the process of user interface translation and validation. Only criteria that were unanimously accepted were included in the definition of language quality. The items of the questionnaire were then derived from the following formal definitions of language quality: friendliness, casualness, professionalism, naturalness, easy-to-understand, appropriateness, correctness and global satisfaction. The final set of items can be found in Table 1.

3.1.2. Scale

To reduce room for interpretation, cultural effects, and translation problems, it was decided to use a 5-point Likert-scale with fully labeled scale points. All scale labels can be found in Table 1.

3.1.3. Experimental procedure

In order to validate the LQS, it was implemented as an online survey and tested with English-speaking users from the US that were recruited on the platform YouTube with an in-product survey link. Participation was voluntary (opt-in) and no compensation was offered for taking part in the study. Users were asked to rate the text quality of the YouTube interface. All 10 items were presented in sequential order. At the end of the survey, users had the opportunity to provide open-text comments on the questionnaire. There were no major redesigns of YouTube during the time of measurement.

No.	Item	Scale
1	How friendly or unfriendly is the text used in the [product name] interface? By "friendly" we mean that the language used shows that [product name] respects and likes their users.	Very unfriendly; Rather unfriendly; Neither unfriendly nor friendly; Rather friendly; Very friendly
2	How casual or formal is the text used in the [product name] interface? By "casual" we mean that the language used is relaxed, like friends speaking to each other. By "formal" we mean that the language is academic, similar to the text of an essay or a legal document.	Very formal; Rather formal; Neither formal nor casual; Rather casual; Very casual
3	How professional is the text used in the [product name] interface? By "professional" we mean that the language is well-written and shows that [product name] cares about quality.	Not at all professional; Slightly professional; Moderately professional; Very professional; Extremely professional
4	How natural or unnatural is the text used in the [product name] interface? Natural here means that the language used represents the way people normally speak to each other.	Very unnatural; Rather unnatural; Neither unnatural nor natural; Rather natural; Very natural
5	How easy or difficult to understand is the text used in the [product name] interface?	Very difficult to understand; Rather difficult to understand; Neither difficult nor easy to understand; Rather easy to understand; Very easy to understand
6	How appropriate or inappropriate do you consider the text in the [product name] interface?	Very inappropriate; Rather inappropriate; Neither inappropriate nor appropriate; Rather appropriate; Very appropriate
7	How often do you encounter grammatical errors in the text used in the [product name] interface?	Always; Often; Sometimes; Rarely; Never
8	How often do you encounter typos / spelling errors in the text used in the [product name] interface?	Always; Often; Sometimes; Rarely; Never
9	How often do you encounter untranslated words that are not in English in the text used in the [product name] interface?	Always; Often; Sometimes; Rarely; Never
10	How satisfied or dissatisfied are you with the quality of language in the [product name] interface when using English?	Very dissatisfied; Rather dissatisfied; Neither dissatisfied nor satisfied; Rather satisfied; Very satisfied

Note. For this study, [product name] was replaced with "YouTube".

Table 1: The first version of the LQS

3.1.4. Sample and data cleaning

A total of 3588 participated in the survey. This sample was subject to a rigorous data cleaning procedure described here:

1. YouTube not only provides linguistic user-interface elements, but also large amounts of user-generated language. The survey instructions clarified that users should only think about user interface elements when answering the survey ("...would like you to think about the written language provided by YouTube in elements such as buttons, information dialogues, navigation or help text, not the text provided within video titles, descriptions, audio tracks or comments."). To control whether users had read and followed this instruction, we asked them at the end of the survey: "Please tell us which of the following text elements came into your mind while rating the language quality of the YouTube interface". With this procedure, a total of 2188 had to be removed because they indicated that they rated the language quality of user-generated content.
2. For this analysis we decided to include only native speakers. Therefore we asked participants to "Rate the level of your reading skills in English" (answers: Basic, Moderate, Fluent, Native). A total of 397 participants who did not choose "Native" were excluded.
3. Another important factor was whether users interact with the user interface often enough to make an accurate judgement of its language quality. Accordingly, a total of 15 participants were excluded because they indicated using YouTube less than once a week.
4. Because we wanted to assess only the English version of YouTube,

people who indicated using YouTube also in non-English languages were removed from analysis. This was the case for 135 participants.

5. Another 13 participants were removed because they could be identified as spam or left more than half of the items unanswered.

3.2. Results

The remaining sample consisted of n = 843 responses. The majority were male (73.5 % male; 19.6 % female; 6.9 % did not indicate their sex) and 55.4 % were between 18 and 29 years old. The gender distribution appeared to be skewed towards the male population. A comparison to the overall YouTube gender distribution was not possible, because there are no exact numbers (a significant amount of YouTube users do not provide their gender or age). The sample's demographic characteristics can be found in Table 2.

Table 3 offers an overview of all missing values for each item. To prevent further sample size reduction with listwise and pairwise deletion, the Expectation-Maximization Algorithm (EM) was used to replace missing values. EM is a valid and reliable method to replace missing values. It is generally preferred over listwise and pairwise deletion (Allison, 2002; Schafer and Graham, 2002) and is often used in survey validation research (Bargas-Avila et al., 2009, 2010).

Table 4 shows the statistics for the first validation. The distribution skewed negatively towards the higher end of the scale, therefore data were log-transformed for further analysis. Transformation is a widely used method to ensure normal distribution of data (Tabachnick and Fidell, 1996). The difficulty indices ranged between .69 and .86, which means that participants tended to answer the items positively.

Sex	N	%	Age	N	%
Female	165	19.6	17 or younger	126	14.9
Male	620	73.5	18 - 29	467	55.4
Not indicated	58	6.9	30 - 39	73	8.7
Total	843	100	40 - 49	37	4.4
			50 - 59	23	2.7
			60 or older	11	1.3
			Not indicated	106	12.6
			Total	843	100

Table 2: Demographics of participants in the first validation

Item	1	2	3	4	5	6	7	8	9	10
N	842	841	842	833	836	835	840	831	827	839
Missing	1	2	1	10	7	8	3	12	16	4
in %	0.1	0.2	0.1	1.2	0.8	0.9	0.4	1.4	1.9	0.5

Table 3: Missing values for each item

According to Fisseni (2004) it is advisable to calculate the discriminatory power with a product-moment correlation of the item score with the test score for interval-scaled item responses. If the items of a scale have moderate to high positive corrected item-total correlations one can expect that the items measure a similar construct as the total score of a questionnaire (Moosbrugger and Kelava, 2007). This means that in case of high discriminatory power, the respondents score for this item reflects the sum score of all other items for this particular respondent. The discriminatory power and Cronbach's α for each item are listed in Table 5. The discriminatory coefficients ranged between .15 and .59 with a mean of .45 (SD = .132). Three

Item	M	SD	S	K	p_v
1	3.74	1.119	-0.827	0.135	.685
2	3.31	0.981	-0.258	-0.406	.578
3	3.65	0.944	-0.551	-0.021	.664
4	3.72	0.983	-0.702	0.180	.682
5	4.24	0.879	-1.263	1.617	.813
6	4.20	0.921	-1.147	1.149	.803
7	4.40	0.860	-1.854	4.061	.851
8	4.44	0.867	-2.027	4.613	.864
9	4.28	0.966	-1.310	1.175	.824
10	4.21	0.929	-1.304	1.750	.805

Note. $N = 843$; Missing values = EM; $SE_S = .084$; $SE_K = .168$;
S = Skewness; K = Kurtosis; p_v = difficulty indices

Table 4: Statistics, first validation (untransformed)

Item	1	2	3	4	5	6	7	8	9
r_{it}	.408	.150	.506	.508	.482	.585	.523	.542	.365
α_{-i}	.751	.787	.733	.733	.738	.722	.732	.729	.755

Note. r_{it} = corrected item - total correlation; α_{-i} = Cronbach's α if item deleted;
$\alpha_{item1-9} = .765$. $N = 843$; Missing values = EM

Table 5: Discriminatory power and Cronbach's α (first version)

items showed a coefficient below .50 (items 1, 2, 5 and 9). According to Borg and Groenen (2005) the lowest acceptable discriminatory power is .30. Item 2 showed a coefficient of .15. The rest of the items were in an acceptable to good range.

Homogeneity examines whether all items of the LQS measure the same construct ("language quality") and whether there are items that overlap (measure similar aspects of the construct). We calculated this by averaging the inter-item correlations for each item (Briggs and Cheek, 1986) similar to the study by Bargas-Avila et al. (2009). The intercorrelation matrix (see Ta-

ble 6) depicts this aspect with significant correlations for all items ($p < .01$) except for item 2 which showed non-significant correlations with items 6, 7, 8, 9 and 10 as well as a significant correlation with item 3 on a higher α-level ($p < .05$). The global item 10 showed moderate correlations in a range of .11 to .48, with all items, with item 2 showing the lowest correlation (.11). The average homogeneity index for the scale was at .28 and the homogeneity indices for each item ranged from .09 to .36 with the lowest value for item 2 (.09). A possible explanation for the relatively moderate indices could be the complexity of the measured construct "language quality", which is composed of many different aspects of language.

Item	1	2	3	4	5	6	7	8	9	10
1	1									
2	.278	1								
3	.348	.080*	1							
4	.319	.219	.321	1						
5	.264	.154	.354	.443	1					
6	.328	.051†	.441	.441	.477	1				
7	.164	$-$.038†	.346	.233	.207	.361	1			
8	.139	$-$.035†	.339	.247	.234	.385	.842	1		
9	.114	$-$.003†	.188	.187	.160	.271	.438	.496	1	
10	.275	.106†	.420	.370	.416	.477	.289	.311	.232	1
H	.248	**.090**	.315	.309	.301	.359	.316	.329	.231	.322

Note. *, $p < .05$; †, n.s.; unmarked correlations are significant ($p < .01$).

Table 6: Intercorrelation matrix and homogeneity indices for item 1 – 10 (first version)

Cronbach's α for the LQS was moderate with .765, suggesting an acceptable reliability for the first version of this questionnaire. Item 10 was not included in the reliability analysis because it reflects a user's global evaluation of the language quality and could artificially inflate Cronbach's α.

Table 5 shows that the internal consistency could be improved if item 2 is excluded.

3.3. Discussion of the first version of the LQS

The first validation of the LQS shows promising results. It also becomes clear that item 2 needs to be modified or deleted.

3.3.1. Scale

There is a tendency to use the LQS in the upper part of the five-point scale. This is not surprising, as YouTube is created and translated by professional linguists and therefore it can be expected that the language quality is rather good. Also this is in line with other research on satisfaction surveys which shows that these items are commonly answered in the upper part of the scales (Bargas-Avila et al., 2009).

3.3.2. Items

Item 2 showed insufficient statistical values in terms of low correlation with other items and unsatisfactory homogeneity index. The reliability of the questionnaire increases after deletion of this item. A closer analysis revealed that the wording "How casual or formal is the text used in the YouTube interface?" combined two aspects that are difficult to interpret. Casualness and formality are highly subjective aspects and might be perceived and judged very differently by different users. The low discriminatory power points at this problem, therefore item 2 was deleted.

The analysis of the open-ended question at the end of the questionnaire also revealed that some users reported encountering text that did not make sense in their opinion. This aspect was not yet covered with the LQS items.

Hence, a new item was introduced for the next iteration, which would allow measuring the occurrence of nonsensical text ("How often do you encounter text that does not make sense?").

4. Second validation

In the revised LQS the item "How casual or formal is the text used in the YouTube interface?" was removed and a new item, "How often do you encounter text that does not make sense?" was added (see Table 7 for a list of all items).

4.0.3. Experimental procedure

In order to validate the second version of the LQS, it was again implemented and tested in the same way the first version was validated.

4.0.4. Sample and data cleaning

A total of 3327 participants completed the survey. The same data cleaning as in the first study was applied. This way, 2161 participants had to be removed because they indicated that they rated the language quality of user-generated content. From the remaining sample, 333 were non-native English speakers, 7 did not use YouTube at least once a week, 95 used YouTube also in non-English languages, and 41 were removed because they could be identified as spam, left more than half of the items unanswered or answered all questions with the same value.

No.	Item	Scale
1	How friendly or unfriendly is the text used in the [product name] interface? By "friendly" we mean that the language used shows that [product name] respects and likes their users.	Very unfriendly; Rather unfriendly; Neither unfriendly nor friendly; Rather friendly; Very friendly
2	How professional is the text used in the [product name] interface? By "professional" we mean that the language is well-written and shows that [product name] cares about quality.	Not at all professional; Slightly professional; Moderately professional; Very professional; Extremely professional
3	How natural or unnatural is the text used in the [product name] interface? Natural here means that the language used represents the way people normally speak to each other.	Very unnatural; Rather unnatural; Neither unnatural nor natural; Rather natural; Very natural
4	How easy or difficult to understand is the text used in the [product name] interface?	Very difficult to understand; Rather difficult to understand; Neither difficult nor easy to understand; Rather easy to understand; Very easy to understand
5	How appropriate or inappropriate do you consider the text in the [product name] interface?	Very inappropriate; Rather inappropriate; Neither inappropriate nor appropriate; Rather appropriate; Very appropriate
6	How often do you encounter grammatical errors in the text used in the [product name] interface?	Always; Often; Sometimes; Rarely; Never
7	How often do you encounter typos / spelling errors in the text used in the [product name] interface?	Always; Often; Sometimes; Rarely; Never
8	**How often do you encounter text that does not make sense in the text used in the [product name] interface?**	**Always; Often; Sometimes; Rarely; Never**
9	How often do you encounter untranslated words that are not in English in the text used in the [product name] interface?	Always; Often; Sometimes; Rarely; Never
10	How satisfied or dissatisfied are you with the quality of language in the [product name] interface when using English?	Very dissatisfied; Rather dissatisfied; Neither dissatisfied nor satisfied; Rather satisfied; Very satisfied

Note. For this study, [product name] was replaced with "YouTube". Item no. 8 (bold) was added for this second version of the LQS.

Table 7: The second version of the LQS

Sex	N	%	Age	N	%
Female	123	17.8	17 or younger	123	17.8
Male	524	75.9	18 - 29	411	59.6
Not indicated	43	6.2	30 - 39	56	8.1
Total	690	100	40 - 49	29	4.2
			50 - 59	11	1.6
			60 or older	4	0.6
			Not indicated	56	8.1
			Total	690	100

Table 8: Demographics of participants in the second validation

4.1. Results

The remaining sample consisted of n = 690 responses. As with the first study, the majority of the participants were male (75.9 % male; 17.8 % female; 6.2 % did not indicate their sex) and 59.6 % were between 18 and 29 years old (see Table 8).

Table 9 provides an overview of all missing values for each item. As described before, the Expectation-Maximization Algorithm (EM) was used to replace the missing values.

Item	1	2	3	4	5	6	7	8	9	10
N	689	689	683	684	684	682	676	678	672	679
Missing	1	1	7	6	6	8	14	12	18	11
in %	0.1	0.1	1	0.9	0.9	1.2	2	1.7	2.6	1.6

Table 9: Missing values for each item (second version)

Table 10 shows the statistics for the second validation. As with the first version, the distribution of the item values skewed negatively towards the higher end of the scale, therefore data were log-transformed for further analysis. The difficulty indices ranged between .65 and .85, which reflects the participants' tendency to answer the items positively.

Item	M	SD	S	K	p_v
1	3.60	1.041	−0.577	−0.049	.651
2	3.64	0.873	−0.429	0.130	.661
3	3.65	0.983	−0.665	0.142	.664
4	4.16	0.869	−0.931	0.595	.791
5	4.11	0.920	−0.895	0.563	.779
6	4.37	0.846	−1.449	2.083	.846
7	4.39	0.852	−1.546	2.336	.853
8	4.14	0.946	−0.989	0.535	.790
9	4.23	0.983	−1.140	0.500	.813
10	4.06	0.936	−0.980	0.907	.770

Note. $N = 690$; Missing values = EM; $SE_S = .093$; $SE_K = .186$;
S = Skewness; K = Kurtosis; p_v = difficulty indices

Table 10: Statistics, second validation (untransformed)

The discriminatory power and Cronbach's α for each item are listed in Table 11. The discriminatory coefficients ranged between .39 and .63 with a mean of .52 ($SD = .085$). Five items showed a coefficient below .50 (item 1, 2, 3, 4 and 9). All items showed satisfactory values.

To explore the homogeneity, the intercorrelation matrix (see Table 12) depicts all significant correlations ($p < .01$). The global item 10 correlated in a range from .29 to .46 with all items, showing low to moderate correlations. The average homogeneity index for the scale is .36 and the homogeneity indices for each item ranged from .26 to .41. Compared to the first version of

Item	1	2	3	4	5	6	7	8	9
r_{it}	.389	.490	.457	.499	.571	.634	.619	.606	.464
α_{-i}	.820	.806	.810	.805	.796	.790	.791	.792	.809

Note. r_{it} = corrected item - total correlation; α_{-i} = Cronbach's α if item deleted; $\alpha_{item1-9}$ = .820. N = 690; Missing values = EM

Table 11: Discriminatory power and Cronbach's α (second version)

the LQS these values show an increase in the intercorrelations for all items.

Item	1	2	3	4	5	6	7	8	9	10
1	1									
2	.402	1								
3	.303	.262	1							
4	.215	.303	.482	1						
5	.326	.445	.436	.497	1					
6	.234	.341	.214	.269	.326	1				
7	.226	.306	.199	.248	.319	.849	1			
8	.209	.282	.310	.392	.330	.577	.580	1		
9	.162	.217	.197	.195	.275	.472	.485	.470	1	
10	.294	.410	.391	.461	.456	.418	.397	.428	.388	1
H	.263	.330	.310	.340	.379	.411	.401	.398	.318	.405

Note. All correlations are significant ($p < .01$).

Table 12: Intercorrelation matrix and homogeneity indices for item 1 – 10 (second version)

Cronbach's α for the LQS was high with .820, suggesting very good reliability for the second version of this questionnaire. In most cases, values for Cronbach's α above .70 are acceptable to good, values between .80 and .90 are very good and values above .90 might indicate item redundancy (DeVellis, 2012). Again, item 10 was excluded from the reliability analysis. Table 11 shows that the internal consistency cannot be improved with the exclusion of any of the items.

4.2. Exploratory factor analysis

In order to investigate the structure of the items, a principal component analysis was conducted. Again, global item 10 was excluded from the analysis. The solution revealed two factors with an eigenvalue greater than 1.00, explaining 58.2 % of the total variance. The factors were rotated using the Oblimin method with Kaizer Normalization. Oblimin rotation was chosen because it is reasonable to expect that the emerging factors are correlated. Analysis showed that the emerging factors correlated with $r = .429$. The factor scores of both factors, calculated with regression method, correlated significantly with the global item 10 ($r_{1(LC)} = .486$; $r_{2(R)} = .564$; $p < 0.001$). The factor loadings for the extracted factors are shown in Table 13. An interpretation based on factor loadings suggests that the first factor describes the frequency of (in)consistencies in the language (Linguistic Correctness) and the second factor describes how natural and smooth to read the used language is (Readability).

In conclusion, the data show evidence that the LQS has a bi-dimensional structure, covering the factors "Linguistic Correctness" (items 6 to 9) and "Readability" (items 1 to 5). The items associated with the two factors can be treated as sub-scales of a global language quality. The scores of the subscales correlate significantly with the global item 10 ($p < 0.01$) with $r = .507$ for linguistic correctness and $r = .573$ for readability. The reliability of the subscales is on a acceptable to good level with $\alpha = .836$ for linguistic correctness and $\alpha = .740$ for readability.

	Factor 1: Linguistic Correctness	Factor 2: Readability
Eigenvalues	3.803	1.440
Friendly (item 1)	0.249	**0.601**
Professional (item 2)	0.374	**0.646**
Natural (item 3)	0.236	**0.738**
Easy to understand (item 4)	0.315	**0.737**
Appropriate (item 5)	0.382	**0.778**
Grammatical errors (item 6)	**0.901**	0.372
Typos / spelling errors (item 7)	**0.907**	0.345
Text does not make sense (item 8)	**0.769**	0.460
Untranslated words (item 9)	**0.707**	0.287

Note. Extraction method: principal component analysis.
Rotation method: Oblimin with Kasier normalization.

Table 13: Exploratory factor analysis

5. Validity and Generalization

5.1. Convergent validity

Convergent validity was examined by exploring the relationship of the LQS with an established measurement of usability. In a study with a final set of n = 211 native English speakers on YouTube (same data cleaning applied as described in prior sections), participants answered the Usability Metric for User Experience (UMUX), before filling out the LQS (second

and final version as described in section 4). UMUX (Finstad, 2010) is a reduced version of the SUS (Brooke, 1996), and contains four items measuring perceived effectiveness, efficiency, satisfaction and overall usability. Finstad showed that UMUX is a reliable, valid and sensitive alternative to SUS if a shorter metric is needed.

The reliability metrics of LQS and UMUX were high (Cronbach's Alpha $\alpha = .829$ and $\alpha = .813$). The correlation of the overall LQS score with the convergent construct "usability" was moderate ($r = .396$, $p < .01$, $N = 211$). The LQS subscale Readability correlated with UMUX on a moderate level ($r = .446$, $p < .01$, $N = 211$), substantially stronger than the subscale Linguistic Correctness ($r = .157$, $p < .05$, $N = 211$).

Conceptually a moderate correlation between two related (but not identical) constructs is to be expected. A very low correlation would hint at the fact that language quality and usability are not correlated or that the LQS does not measure the targeted construct. A very high correlation would mean that both constructs overlap strongly and would question the necessity of a separate survey. In the case of LQS, the moderate correlations are evidence that it measures a construct that relates to usability, but is different enough to warrant a separate survey.

A possible explanation for the different correlation strengths could be that Readability contains aspects of language that are more directly related to usability. For instance ease of understanding or naturalness of the UI text might have a direct impact on product usability. In contrast, Linguistic Correctness, which describes aspects like typos or grammar errors, seems to impact ease of use less strongly.

These data provide evidence for convergent validity of the LQS. Language quality and usability are constructs that partly overlap, but are not the same. Language quality can't be regarded as a stand alone aspect of a user interface – it clearly correlates with usability ratings, though on a moderate levels.

5.2. Discriminative validity

To further examine the validity of the LQS, discriminative validity was examined. During data cleaning (see section 4.0.4), all participants who indicated to have rated user-generated content (video titles, video descriptions or audio tracks) were removed from the analysis. To calculate discriminative validity, these data were used. It is reasonable to assume, that user generated language (UGL) will be of less quality than the expert-generated language of the YouTube user interface.

A sample of 430 participants who rated only UGL was identified. The average score (items 1 to 9) of this group is 3.36 ($SD = .756$). These levels are significantly lower than the score for the YouTube user interface ($\bar{x} = 4.03, SD = .593, N = 690$), $t(752.184) = 15.645, p < .001, d = 0.99$ (large effect).

This analysis provides further evidence, that the LQS is a valid tool to measure language quality. Participants who rated user-generated language, provided significantly lower scores than users rating language that was created by experts.

5.3. Generalization to other languages

A key question of the LQS was: Would it scale to other languages and deliver valuable data? To answer this question, the LQS needed to be trans-

lated into other languages and new data had to be gathered.

To do this, the survey was translated for a selection of languages that show high YouTube usage. The survey was first translated by a professional linguist and then reviewed by a second one. Both translators received detailed instructions on aspects they should pay attention to. All parts that led to disagreement were discussed and resolved between the translators.

Table 14 shows a summary of the key statistics. The numbers show similar values for all languages. While the sample sizes vary, the number of missing data points is comparable and relatively low for each language. Similar to the English version, item difficulties tend towards the higher end, which is probably due to the relatively high quality of the YouTube user interface language. The discriminatory power is satisfactory but some items were below the recommended value of .3 for Portuguese-Brazil and French. The homogeneity of the items in other languages is – similar to the English version – on the lower end and reflects the relatively complex construct of language quality. The values of Cronbach's α range between .755 and .849 which is an acceptable to good level.

Overall, the validation revealed that the translated versions of the LQS worked as expected and can be applied to measure user interface language quality.

5.4. Generalization to other products

To understand if the LQS can be generalized to other products than YouTube, we ran this study for two entirely different products: Google Analytics and Google AdWords. Analytics is a tool that allows website owners to track and understand their website traffic, AdWords is the platform that

	N N	% mis (min)	% mis (max)	p_v (min)	p_v (max)	r_{it} (min)	r_{it} (max)	H (min)	H (max)	α_{1-9}
English (USA)	690	0.1	2.6	.651	.840	.389	.634	.263	.411	.820
French (France)	308	1.9	5.2	.660	.870	.305	.593	.201	.377	.766
German (Germany)	1016	0.6	2.5	.640	.850	.342	.554	.221	.329	.774
Italian (Italy)	896	0.2	3.2	.690	.870	.329	.597	.217	.359	.793
Portuguese-BR (Brazil)	410	0.7	5.4	.640	.850	.241	.592	.276	.340	.774
Russian (Russia)	358	0.6	3.4	.730	.920	.406	.548	.253	.347	.781
Spanish (Spain)	333	0.9	3.3	.610	.840	.451	.615	.274	.381	.825
Spanish LatAm (Mexico)	300	0	2.3	.640	.830	.429	.620	.310	.423	.844
Hebrew (Israel)	178	1.8	3.4	.669	.890	.379	.643	.260	.414	.828
Arabic (Saudi Arabia, Egypt, UAE, Morocco)	95	1.1	8.4	.580	.850	.394	.707	.270	.463	.849

Note. mis = missing values; p_v = item difficulty; r_{it} = discriminatory power; H = homogeneity; α = internal consistency

Table 14: Statistics of the LQS in other languages

allows advertisers worldwide to buy, configure and track advertisement that is run on Google properties. If the LQS is product independent, key statistics should be similar, no matter if the surveys are answered by consumers (YouTube), website owners (Analytics) or advertisers (AdWords).

	N	p_v (min)	p_v (max)	r_{it} (min)	r_{it} (max)	H (min)	H (max)	α_{1-9}
YouTube *	690	.651	.840	.389	.634	.263	.411	.820
Google Analytics *	902	.580	.880	.360	.616	.257	.431	.811
Google AdWords *	400	.670	.900	.368	.632	.249	.386	.809

Note. p_v = item difficulty; r_{it} = discriminatory power; H = homogeneity; α = internal consistency

* shown values are for LQS in English

Table 15: Generalization of LQS to other products

The item analysis for these two additional products revealed key statistic values that are close to the results for the YouTube Interface (see Table 15),

providing evidence that the LQS can indeed be generalized to other products.

6. Case Study: Applying the LQS in the field

The main reason for developing the LQS was to discover problematic translations of the YouTube interface to allow focused quality improvement efforts. To do this, the LQS was translated to over 60 languages and data were gathered for all these versions of the YouTube interface. While the exact results for each language are not the topic of this paper, a high level overview of the process and results are provided to practitioners:

- To understand quality of each UI version, we compared the results for the translated versions to the source language (here: English). We inspected first the global item, in combination with Linguistic Correctness and Readability. No further weighting was applied. Second, we inspected each item separately, to understand which notion of Linguistic Correctness or Readability showed worse (or better) values.

- The data revealed that about one third of the languages showed subpar language quality levels, when compared to the source language

- To understand the source of these problems and fix them, two actions were taken: (1) run a modified version of the LQS to gather qualitative feedback, and (2) conduct in-depth quality reviews with experts (as recommended by Schriver, 1989)

- The modified version of the LQS consisted of the identical survey, with one slight change. Every time a survey respondent selected the lower

two end scale points, pointing to a problem in the language, a text box with the following question was surfaced: "Can you tell us what to improve? Any examples or links would help us understand what needs to be changed.". With this approach we aimed at generating more actionable qualitative knowledge on how to improve translations.

The analysis of these comments provided linguists with valuable feedback of various kinds. For instance, users pointed to confusing terminology, untranslated words that were missed during translation, typographical or grammatical problems, words that were translated but are commonly used in English, or screenshots in help pages that were in English but needed to be localized. Some users also pointed to readability aspects such as sections with old fashioned or too formal tone as well as too informal translations, complex technical or legal wordings, unnatural translations or rather lengthy sections of text. In some languages users also pointed to text that was too small or criticized the readability of the font that was used. Experts did not always agree with the qualitative feedback from users. Many comments triggered fruitful conversations, of which not all led to changes.

- In parallel, in-depth expert reviews (so called "language find-its") were organized. In these sessions, a group of experts for each language met and screened all of YouTube to discover aspects of the language that could be improved. All problems were gathered, discussed in the team, and concrete actions decided on how to fix them. By using the LQS data to select the target languages, it was possible to reduce the number of language find-its to about one third of the original estimation (if all

languages had been screened).

In summary it can be said that the LQS proved a reliable, valid and useful tool to approach language quality evaluation and improvement.

7. Discussion

7.1. Summary and conclusions

There are three approaches to evaluate the quality of text (Schriver, 1989). (1) Text-based evaluation methods such as automated readability scores can be easily calculated and are usually cost-effective, but their usefulness for improving language is rather superficial. (2) Expert-judgement based methods create in-depth actionable insights, but these approaches are limited due to the lack of an outside perspective, their difficulty in anticipating text problems on a user level and the high costs associated with them. (3) Reader-focused methods can be quite cost-efficient, provide user-centric perspectives, but generate few actionable insights on how to improve the language. Therefore, a combination of expert-judgement and reader-focused methods is promising.

This article presents the development and validation of a reader-focused method: A survey enables companies to have their users rate the language quality provided in the user interface. Professional linguists agreed upon central aspects of language quality. Based on this, 10 items for the first version of the LQS were developed. This questionnaire was applied in an online survey in order to evaluate user interface text quality and to validate the questionnaire. The item analysis of the first version of the LQS revealed that one item did not satisfy statistical criteria and therefore was eliminated

from the tool. After the first validation, qualitative user feedback suggested that the inclusion of an item to cover the occurrence of nonsensical text in the user interface would help users in the rating process. A new question was added to the questionnaire to measure this aspect.

The second version of the questionnaire showed good statistics. An exploratory factor analysis revealed that the questionnaire measures two factors: 1) more objective aspects such as typography, grammar and frequency of untranslated words that were summarized under the term *Linguistic Correctness* and 2) rather subjective aspects such as friendliness and appropriateness, named *Readability*.

Both validations of the LQS showed high Cronbach's α levels, which is clear evidence of good internal consistency. The second validation indicates that Cronbach's α cannot be increased further by the exclusion of any item. For the second validation, the homogeneity indices have been increased to an acceptable level. Given the complexity of the construct "language qualtiy", heterogeneous items can be expected. Thus, the overall reliability and validity of the LQS are good. Good content validity can be assumed, as all items were developed and approved by a group of expert linguists, making it very likely that the most important aspects of language quality have been considered. Criterion-related validity is measured by the correlations with the global item, which also showed satisfactory results. There is clear evidence for convergent validity, as shown by the correlation to UMUX, as well as discriminative validity, as shown by the analysis of user-generated vs expert-generated content. There is evidence that the validation of the LQS might be language-independent, because the analysis of other languages showed sim-

ilar results. This survey can also be generalized to other products, because the application to Google Analytics and Google AdWords revealed similar survey validation statistics.

While it can be criticized that the questionnaire at hand measures language quality retrospectively, a concurrent reader-focused measure for the user interface language quality of a global website is not feasible and would be extremely expensive to accomplish. In general, the vast majority of questionnaires in the field of usability are applied post-use (Hornbaek, 2006).

In order to reach users worldwide, localization and translation is important. Even seemingly small differences such as having an Australian English version of a website as opposed to an international English version can make a difference for users: "And even in English-speaking Australia, users strongly preferred local sites to foreign sites. Although they could read both American and English-language European sites just fine, Aussie users felt that foreign sites weren't as relevant to their needs." (Nielsen, 2011). While many other aspects of design such as color use, symbols and icons, as well as technical aspects such as date and time formats are important, a lot of the information is also conveyed through text.

Del Galdo and Nielsen (1996) argue that there are three levels at which to tackle the problem of producing international user interfaces. The first level is the technical implementation of users' native language character set, notation and formats. This can be regarded as accomplished by most companies, according to del Galdo and Nielsen. The second level is producing a user interface and user information that are understandable and usable in the user's native language. The LQS aims to help reach this level by providing

user-feedback about linguistic correctness and readability in order to asses and improve the text quality of a user interface. This is the foundation of the third level, proposed by del Galdo and Nielsen: the ability to produce systems that accommodate cultural characteristics of the users. This means that designs must address specific cultural models, such as the way people communicate or the way business is conducted in different countries.

The LQS allows practitioners to identify translations that need quality improvement which in turn allows the efficient allocation of resources to conduct expert-judgement based reviews. Also, the questionnaire can be applied at different stages of the product to measure the effect of changes. It is beneficial to combine the evaluation metrics with qualitative feedback. Allowing participants to provide reasons for their low rating on certain items has been proven to be useful for the derivation of actionable in-sights. The LQS has been extensively tested in the evaluation of the YouTube UI translation quality and helped to improve the language quality and ultimately the quality of the user experience.

Perceived language quality of translated user interfaces can have a significant impact on the perception of the overall quality and usability of a product. It is therefore important to assess and improve the quality of language used in applications. The LQS can be regarded as a small piece in the puzzle of understanding and improving language quality.

7.2. Limitations

There are several limitations of this study: (1) Similar to most survey based approaches, participation was "opt-in". This means that respondents could chose if they answer or not, which can lead to sampling biases. While

this problem is present for almost all survey based approaches, it is important to keep in mind when interpreting results. (2) In this publication, the LQS was applied only to browser based websites on desktop computers. Additional studies are needed to understand if it can be generalized to other applications, such as for instance mobile apps. (3) As stated before, there are several approaches to measure language quality. The LQS allows only a subjective user-based post-usage measurement and needs to be combined with other methods to deliver the full picture.

7.3. Future research

Future research could increase the validity of the survey by comparing post- to pre-revision results of the LQS. Practitioners and researchers might also benefit from a benchmark, which provides industry standards for good and bad LQS values. Another step could be to develop and validate a short version of the LQS that would allow measuring UI text quality in mobile context/applications.

8. Acknowledgements

We would like to thank the following people for their great help and support bringing this project to life: Devesh Kothari, Fredrik Lundh, Günther Noack, Keumhee Jeong, Matthew Glotzbach, Meike Schmidt, Olga Khroustaleva, Oliver Heckmann, Patricia Gómez Jurado, Svein Hermansen, and Wojtek Cyprys. Also, we would like to thank Sebastien Orsini for his help in survey validation methods, and Alexandre Tuch and Kasper Hornbæk for giving us feedback on early versions of the manuscript.

9. References

Allison, P. D., 2002. Missing data: Quantitative applications in the social sciences. British Journal of Mathematical and Statistical Psychology 55 (1), 193–196.

Bargas-Avila, J. A., Lötscher, J., Orsini, S., Opwis, K., Nov. 2009. Intranet satisfaction questionnaire: Development and validationof a questionnaire to measure user satisfaction with the intranet. Computers in Human Behavior 25 (6), 1241–1250.
URL http://dx.doi.org/10.1016/j.chb.2009.05.014

Bargas-Avila, J. A., Orsini, S., de Vito, M., Opwis, K., Jan. 2010. Zego: Development and validation of a short questionnaire to measure user satisfaction with e-government portals. Advances in Human-Computer Interaction 2010, 6:1–6:10.
URL http://dx.doi.org/10.1155/2010/487163

Blackmon, M. H., Kitajima, M., Polson, P. G., 2005. Tool for accurately predicting website navigation problems, non-problems, problem severity, and effectiveness of repairs. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '05. ACM, New York, NY, USA, pp. 31–40.
URL http://doi.acm.org/10.1145/1054972.1054978

Borg, I., Groenen, P. J., 2005. Modern multidimensional scaling: Theory and applications. Springer.

Briggs, S. R., Cheek, J. M., 1986. The role of factor analysis in the development and evaluation of personality scales. Journal of Personality 54 (1), 106–148.
URL http://dx.doi.org/10.1111/j.1467-6494.1986.tb00391.x

Brooke, J., 1996. Sus-a quick and dirty usability scale. In: Usability evaluation in industry. Vol. 189. London: Taylor & Francis, p. 194.

Chi, E. H., Pirolli, P., Chen, K., Pitkow, J., 2001. Using information scent to model user information needs and actions and the web. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '01. ACM, New York, NY, USA, pp. 490–497.
URL http://doi.acm.org/10.1145/365024.365325

Del Galdo, E., 1990. Internationalization and translation: Some guidelines for the design of human-computer interfaces. In: Nielsen, J. (Ed.), Designing User Interfaces for International Use. Elsevier Science Publishers Ltd., Essex, UK, pp. 1–10.
URL http://dl.acm.org/citation.cfm?id=130347.132705

Del Galdo, E. M., Nielsen, J., 1996. International users interface. John Wiley & Sons, Inc.

DeVellis, R. F., 2012. Scale development: Theory and applications, 3rd Edition. Vol. 26 of Applied Social Research Methods Series. Sage Publications.

Dilts, D. W., 2001. Successfully crossing the language translation divide. In: Proceedings of the 19th Annual International Conference on Computer

Documentation. SIGDOC '01. ACM, New York, NY, USA, pp. 73–77.
URL http://doi.acm.org/10.1145/501516.501531

Finstad, K., 2010. The usability metric for user experience. Interacting with Computers 22 (5), 323–327.

Fisseni, H. J., 2004. Lehrbuch der psychologischen Diagnostik: mit Hinweisen zur Intervention. Hogrefe Verlag.

Fry, E., 1968. A readability formula that saves time. Journal of Reading 11 (7), 513–578.

Hornbaek, K., 2006. Current practice in measuring usability: Challenges to usability studies and research. International Journal of Human-Computer Studies 64, 79–102.

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., Chissom, B. S., 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Tech. rep., DTIC Document.

Leiva, L. A., Alabau, V., 2014. The impact of visual contextualization on ui localization. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '14. ACM, New York, NY, USA, pp. 3739–3742.
URL http://doi.acm.org/10.1145/2556288.2556982

Lentz, L., de Jong, M., Sep 1997. The evaluation of text quality: expert-focused and reader-focused methods compared. Professional Communication, IEEE Transactions on 40 (3), 224–234.

Lewis, M. P., Simons, G. F., Fennig, C. D. (Eds.), June 2013. Ethnologue: Languages of the World, 17th Edition. Dallas: SIL International.
URL www.ethnologue.com

Moosbrugger, H., Kelava, A., 2007. Testtheorie und Fragebogenkonstruktion. Springer.

Muntés Mulero, V., Paladini Adell, P., España Bonet, C., Màrquez Villodre, L., et al., 2012. Context-aware machine translation for software localization. In: Proceedings of the 16th Annual Conference of the European Association for Machine Translation. EAMT 2012. EAMT, Trento, Italy, pp. 77–80.

Nielsen, J., 1990. Usability testing of international interfaces. In: Nielsen, J. (Ed.), Designing User Interfaces for International Use. Elsevier Science Publishers Ltd., Essex, UK, pp. 39–44.
URL http://dl.acm.org/citation.cfm?id=130347.132707

Nielsen, J., June 2011. International usability: Big stuff the same, details differ. http://www.nngroup.com/articles/international-usability-details-differ/.

Ransdell, S., Levy, C. M., 1996. Working memory constraints on writing quality and fluency. In: Levy, C. M., Ransdell, S. (Eds.), The science of writing: Theories, methods, individual differences, and applications. Lawrence Erlbaum Associates, Inc, pp. 93–105.

Schafer, J. L., Graham, J. W., 2002. Missing data: our view of the state of the art. Psychological Methods 7 (2), 147.

Schriver, K. A., 1989. Evaluating text quality: The continuum from text-focused to reader-focused methods. Professional Communication, IEEE Transactions on 32 (4), 238–255.

Sun, H., 2001. Building a culturally-competent corporate web site: An exploratory study of cultural markers in multilingual web design. In: Proceedings of the 19th Annual International Conference on Computer Documentation. SIGDOC '01. ACM, New York, NY, USA, pp. 95–102. URL http://doi.acm.org/10.1145/501516.501536

Tabachnick, B., Fidell, L., 1996. Using Multivariate Statistics, 3rd Edition. Harper Collins College Publishers, New York.