Usabilidade de um Aplicativo de Libras para Educação e Governança em uma Cidade Amazônica

Ana Paula Martins da Costa[1], Eduardo de Medeiros Diniz[1], Erick Gaia Sales[1], Felipe Coelho[2], Walber Christiano Lima da Costa[3], Léia S. de Sousa[1]

[1]Faculdade de Sistemas de Informação (FACSI)
Universidade Federal do Sul e Sudeste do Pará (UNIFESSPA)
– Marabá – PA – Brasil

[2]Faculdade de Engenharia da Computação (FAEC)
Universidade Federal do Sul e Sudeste do Pará (UNIFESSPA)
– Marabá – PA – Brasil

[3]Faculdade de Ciências da Educação (FACED)
Universidade Federal do Sul e Sudeste do Pará (UNIFESSPA)
– Marabá – PA – Brasil

{nedel,flavio}@inf.ufrgs.br, R.Bordini@durham.ac.uk, jomi@inf.furb.br

Abstract. *This work presents the Pará-LIBRAS mobile application, which is a bilingual Portuguese-LIBRAS glossary, capable of receiving and gathering suggestions from users about new signs, and which intends to be used in the sphere of state, federal education and the legislative chamber of an Amazonian city. The usability of Pará-LIBRAS is evaluated using the Goal-Question-Metric (GQM) and the System Usability Scale (SUS) applied to potential end users. To analyze the validity and reliability of the GQM, it was used considering average variance extracted (AVE), the composite reliability measure (CR), Content Validity Index (CVI), Cronbach1s Alpha, Omega's Coefficient and Raju's consistency. The results of the GQM and SUS demonstrate that positive impressions regarding simplicity, accuracy, time spent, security and attractiveness represent more than 90%.*

Resumo. *Este trabalho apresenta o aplicativo móvel Pará-LIBRAS, que é um glossário bilíngue português-LIBRAS, capaz de receber e reunir sugestões dos usuários sobre novos símbolos, e que pretende ser utilizado na esfera da educação estadual, federal e câmara legislativa de uma cidade amazônica. Avalia-se a usabilidade do Pará-LIBRAS por meio do Goal-Question-Metric (GQM) e do System Usability Scale (SUS) aplicados aos potenciais usuários finais. Para analisar a validade e confiabilidade do GQM considerou-se a Variância Média Extraída (VME), a medida de Confiabilidade Composta (CC), Índice de Validade de Conteúdo (IVC), Alfa de Cronbach, Coeficiente Ômega e Coeficiente de Consistência de Raju. Os resultados do GQM e do SUS demonstram que as impressões positivas em relação a simplicidade, precisão, tempo gasto, segurança e atratividade representam mais de 90%.*

1. Introdução

A ausência de sinais LIBRAS para nomes de pontos turísticos, bairros, vilas e cidades, bem como para identificação e acesso de serviços públicos, é um desafio significativo

para a inclusão e acessibilidade de pessoas surdas em muitas cidades brasileiras. A língua de sinais pode facilitar a aprendizagem de uma língua escrita ou falada, mas também pode haver interferências devido à diferença entre a estrutura visual-gestual e a estrutura auditivo-oral das línguas [Gesser 2012]. Esse é um dos problemas que afetam a sensação de pertencimento à comunidade, identificado em conjunto pela Escola do Legislativo de Marabá, pela Câmara Legislativa e um Centro de Atendimento Especializado na área da Surdez (CAES) de Marabá [Haôr, Victor 2023], uma cidade da região amazônica, especialmente no que se refere ao contexto da identidade das comunidades surdas dentre os povos tradicionais e ribeirinhos. As barreiras comunicacionais ampliaram ainda mais as distâncias geográficas já existentes entre entre esses povos e a sua cidadania [Câmara Municipal de Marabá (CMM) 2024].

As mesmas dificuldades são percebidas também entre docentes regentes, tradutores e intérpretes de LIBRAS que atuam no ensino médio técnico e ensino superior de instituições federais, estaduais e municipais de Marabá. A implementação de metodologias ativas e alinhamento dos conteúdos com a realidade local esbarrou na falta de símbolos para termos técnicos específicos de uma determinada área do conhecimento, os quais geralmente não estão presentes no diálogo entre o professor e aluno surdo, pois só aconteceria de forma plena quando o primeiro adquire a LIBRAS como L1 ou o segundo obtém o português como L2 [Galvão et al. 2020].

Neste contexto surgiu o projeto de produção de um glossário regionalístico/municipal bilíngue português-LIBRAS, onde especialistas em LIBRAS, tradutores/intérpretes e pessoas surdas, provenientes de, ao menos, quatro instituições públicas locais, elaboraram em conjunto expressões gestual-visual para representar e dar sentido ao nome dos bairros, das praças, dos pontos turísticos da cidade e símbolos da cultura local, geralmente vocábulos com origem indígena, entre outros. O conteúdo produzido é veiculado por meio do aplicativo Pará-LIBRAS, proposto neste trabalho.

O Pará-LIBRAS é um aplicativo para dispositivos móveis organizado em diversas categorias de termos, como por exemplo, expressões e cumprimentos regionais, termos matemáticos, termos jurídicos mais comuns utilizados na câmara legislativa, nomes dos bairros, vilas, rios e pontos turísticos da cidade de Marabá. O aplicativo possibilita ao usuário a inclusão de novos termos ou sinais ao glossário existente, bem como a criação de novos categorias de termos. Também é possível editar as categorias existentes. A curadoria dos conteúdos adicionados como contribuição, pode ser realizada pelos administradores do sistema, que podem ser pessoas com surdez e profissionais em português-LIBRAS. O aplicativo encontra-se em fase teste e será utilizado para apoiar o ensino em cursos do ensino médio e graduação, bem como para apoiar a acessibilidade de pessoas com deficiência auditiva aos conteúdos de discussões em sessões da Câmara de Vereadores.

Alinhado com a Trilha 3 (Fatores Humanos em Tecnologia Digital para a Educação) do SBIE 2024, e pertinente aos tópicos *i*) *design* e avaliação de Interfaces de *hardware* e/ou *software* para educação especial, e *ii*) avaliação de acessibilidade, comunicabilidade, experiência do usuário e/ou usabilidade de tecnologias educacionais, este trabalho apresenta resultados e interpretações da avaliação de usabilidade do Pará-LIBRAS, realizado com o público-alvo, com o emprego de duas abordagens: *Goal-Question-Metric* (GQM)[Rombach et al. 1994][Hussain and Kutar 2009] e *System Usability Scale* (SUS)

[Brooke et al. 1996][Lourenço et al. 2022]. Enquanto o GQM é personalizável e envolve uma análise detalhada para cada contexto específico do aplicativo a ser avaliado, e por isso, requer mais tempo e esforço de análise, o SUS é um teste rápido com dez itens fixos, proporcionando uma visão imediata da usabilidade do sistema [Brooke et al. 1996].

Devido a implementação do GQM com questões específicas do contexto do aplicativo, e com diferentes número de questões por objetivo avaliado, foi necessário analisar a validade e confiabilidade do instrumento utilizando as medidas estatísticas variância média extraída (VME), a medida de confiabilidade composta (CC)[Valentini and Damasio 2016], Índice de Validade de Conteúdo (IVC)[Cígler and Chvojka 2022], Alfa de *Cronbach*[Warrens 2015], Coeficiente Ômega e Coeficiente de Consistência de Raju [Raju 1977]. Os resultados de ambos os testes poderão deixar mais clara e objetiva a perspectiva do usuário final em relação ao aplicativo.

Os principais resultados obtidos com a aplicação dos testes demostram que, além de comprovada a validade e confiabilidade do GQM, tanto esse quanto o SUS convergiram na alta taxa de aceitação do aplicativo, com relevância de mais de 90% comprovada em relação às características de usabilidade consideradas fundamentais, como a simplicidade, precisão tempo gasto, segurança e atratividade [do Nascimento 2019][Hussain and Kutar 2009].

2. Revisão de Literatura

Os glossários em Libras são instrumentos fundamentais para promover a inclusão, acessibilidade, e valorização cultural dos surdos, tradutores e intérpretes em diversas áreas do conhecimento. Exemplos de glossários específicos são apresentados em [Friedrich 2019], [Pereira 2021], [Ouedraogo et al. 2020] e [Oliveira et al. 2020].

Um glossário em Libras para o curso de Administração da Universidade Federal de Pelotas, totalizando 102 sinais-termos de 25 palavras, foi proposto em [Friedrich 2019]. Juntamente dos termos representados foram adicionados ainda uma ficha léxica, classe gramatical, exemplos de utilização em uma frase, foto, vídeo e Escrita de Sinais (*SignWriting*). O resultado final foi disponibilizado por meio de um *site* na internet[1]. Um glossário bilíngue de Ortodontia é apresentado em [Pereira 2021]. para cada termo de entrada do referido glossário foi mostrada a categoria gramatical, definição do termo, fonte e exemplo de contexto de utilização do termo. Os termos também foram representados em vídeos no *YouTube*.

Já em [Ouedraogo et al. 2020] apresenta-se um glossário de termos jurídicos em LIBRAS. O resultado final foi apresentado em linguagem audiovisual, tendo sido produzidos um glossário com setenta e quatro sinais-termo, vídeos explicativos de três sinais e dois episódios cinematográficos.

O *software ContaKg* Bilíngue[Oliveira et al. 2020], com termos de Probabilidade e Estatística é voltado para estudantes surdos e ouvintes do ensino fundamental. O *ContaKg* pode ser acessado na web[2]. Foram utilizados GIFs para visualização das informações textuais em Libras. O conteúdo do *software* é um conjunto de três atividades visando a aprendizagem de análise de gráficos simples. As atividades seguem uma

[1]Disponível em `https://www.glossario.libras.ufsc.br/`
[2]Disponível em `https://softwareducativo.github.io/Contakg/`.

narrativa baseada na educação alimentar.

Embora esses trabalhos tenham trazido contribuições com símbolos em Libras para suas respectivas áreas do conhecimento, uma avaliação de usabilidade com o usuário final está ausente em [Friedrich 2019], [Pereira 2021], [Ouedraogo et al. 2020] e [Oliveira et al. 2020], assim como não fica clara a perspectiva da inclusão de novos símbolos pelos usuários dessas ferramentas. O principal diferencial da proposta do atual trabalho é a elaboração do aplicativo juntamente com a comunidade especialista, com surdos, tradutores e intérpretes, bem como a realização da avaliação da usabilidade com dois distintos tipos de testes (GQM e SUS).

Em geral, a metodologia GQM é utilizado para medir uma ampla gama de aspectos do desenvolvimento e manutenção de *software*, não se limitando apenas à usabilidade [Rombach et al. 1994] [do Nascimento 2019]. Os dados coletados com o GQM podem ser objetivos ou subjetivos, dependendo dos mecanismos de medição (método de avaliação quantitativa ou qualitativa). O modelo GQM abrange o processo de determinar o que medir e como medir [Rombach et al. 1994]. Em [McGinn et al. 2018] o foco é a abordagem GQM em sistemas robóticos complexos, cujos resultados contribuíram para melhorias dos processos da tarefa de desenvolver uma garra robótica personalizada para um robô de serviço. Os autores em [Sim et al. 2022] utilizaram o GQM como uma ferramenta interdisciplinar para avaliar o protótipo de um aplicativo móvel de mandarim e também analisaram a confiabilidade do questionário utilizando Alfa de *Cronbach*. Em [Hussain and Kutar 2009] é realizada uma revisão das métricas existentes para aplicações em *desktops* e, com base nas quais apresenta-se um modelo conceitual GQM para avaliar aplicativos móveis, assim como uma discussão sobre como elaborar as perguntas e métricas capazes de avaliar aplicativos.

Já o método SUS, por sua vez, é focado exclusivamente na usabilidade de sistemas e oferece uma maneira rápida e padronizada de avaliar a usabilidade de um sistema [Brooke et al. 1996]. Em [Vlachogianni and Tselios 2022] os autores realizaram uma revisão sistemática de trabalhos no qual o SUS é empregado para avaliar tecnologias educacionais, tais como plataformas de *internet* e *web sites* universitários. Já em [Kaya et al. 2019] investiga-se a diferença, em termos de usabilidade, entre os sistemas operacionais *i*OS e *Android*, quando se trata de aplicativos como *WhatsApp*, *Facebook* e *YouTube*. Em ambos os estudos, os *scores* resultantes foram compatíveis em os resultados de outros tipos de testes de usabilidade. O presente estudo também implementa o SUS e mostra a compatibilidade do resultado com outros estudos.

O GQM e o SUS são simultaneamente empregados em [Stedile et al. 2019] para avaliar a usabilidade de um ambiente de realidade virtual imersivo que auxilia o ensino em conteúdos de oratória. Ambos foram empregados a diversos *frameworks* de *design* de usabilidade para reprojetar a tela de relatório de uma aplicativo móvel sobre terremotos. Os resultados dos testes realizados ajudou a equipe a selecionar o *framework* ideal para a equipe interdisciplinar envolvida. Este trabalho também aplicou os testes GQM e SUS para obter resultados quantitativos e comparáveis das impressões dos participantes.

3. O Aplicativo Pará-LIBRAS

Para a criação do aplicativo com glossário em Libras e símbolos regionalistas/locais, inicialmente realizou-se uma rodada de entrevistas com professores e professoras pesquisa-

dores que atuam em uma universidade federal (UF), instituto federal (IF) e universidade estadual (UE), esquematizado na Figura 1. Com essa coleta de informações, foi notada a necessidade de centralizar os sinais regionais de Libras, incluindo expressões linguísticas utilizadas no estado e outros vocabulários comuns no cotidiano dos entrevistados.

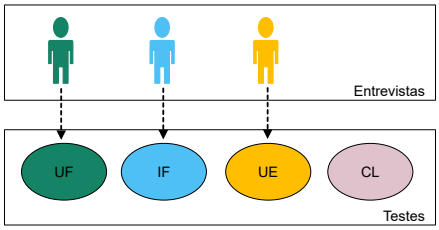

Figura 1. Colabores e participantes da pesquisa

Ao mesmo tempo, profissionais tradutores e intérpretes da Câmara Legislativa Municipal (CL) já haviam iniciado a produção de um glossário com os nomes de alguns projetos de lei de autoria dos vereadores. A partir dos interesses comuns, o aplicativo Pará-LIBRAS foi desenvolvido como uma ferramenta simples e funcional para auxiliar na busca de palavras em Libras, conforme discorre-se nesta seção. Na sequência, testes de usabilidade do Pará-LIBRAS foram aplicados aos participantes entrevistados e seus colegas (Figura 1), nas suas instituições de atuação, conforme será destacado na próxima seção.

Para a construção do Pará-LIBRAS, foram definidas várias etapas. A partir dos dados coletados nas entrevistas, foram extraídos requisitos funcionais e não funcionais. Em seguida, elaborou-se um protótipo visual de baixa fidelidade, validado com os entrevistados. Com o *design* UX definido, iniciou-se a fase de codificação do aplicativo. Utilizou-se *Node.js* por sua eficiência e escalabilidade, *React Native* com *TypeScript* para desenvolvimento multiplataforma, permitindo um único código para *iOS* e *Android*, e *Expo*, para simplificar o desenvolvimento e a implantação, oferecendo serviços e bibliotecas úteis.

O aplicativo foi concebido com dois tipos de usuários: o usuário administrador e o usuário comum. O usuário administrador é responsável por gerenciar o sistema, incluindo a adição e edição de palavras e categorias, bem como a verificação de sugestões para garantir a padronização do sistema. O usuário comum é qualquer indivíduo que utiliza o sistema para consultar sinais em Libras, podendo enviar sugestões de sinais.

A Figura 2A apresenta a tela inicial,onde são exibidas diversos categorias de vocábulos/palavras. Uma categoria de palavras é um grupo de palavras sobre um mesmo tema. São exemplos de “Categorias”o Alfabeto Manual e Expressões Regionais, que agrupam os sinais em Libras com esses temas específicos. Outra forma de acessar esses vocabulários é através do menu (Figura 2B), que centraliza as demais funcionalidades do Glossário. Para os gerenciadores do sistema, são acrescentadas funcionalidades adicionais para gerenciar palavras, categorias, mostradas no quadro em verde na Figura 2B. A categoria “Saudações”é um exemplo de categoria com vários expressões e imagens associadas (Figura 2C) .

Na tela de Gerenciamento de Palavras (Figura 2D), é possível adicionar uma nova

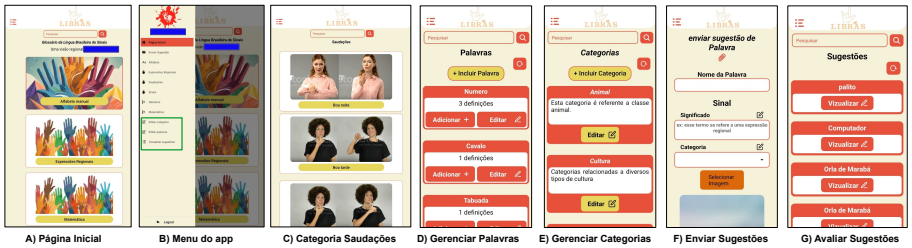

Figura 2. Apresentação das principais telas do aplicativo móvel

palavra ao glossário, adicionar ou editar um sinal para uma palavra já existente, além de visualizar quantas variações de sinais existem para cada palavra no sistema. Na tela de Gerenciamento de Categorias (Figura 2E), é possível adicionar uma nova categoria ou editar uma já existente, além de visualizar a descrição de cada categoria, facilitando a definição dos temas de cada item.

Além disso, considerando a necessidade de centralizar a maior quantidade possível de sinais em Libras, o aplicativo foi concebido com uma segunda funcionalidade principal: permitir que os usuários comuns enviem sugestões de novas palavras e sinais (Figura 2F). Todas as sugestões recebidas serão avaliadas pelos administradores do aplicativo (Figura 2G). Após a validação, feita por um especialista, o sinal recebido deverá seguir o fluxo de padronização e posteriormente ser integrado ao sistema.

Pensando nessas funcionalidades, o aplicativo Pará-LIBRAS se diferencia dos demais por seu objetivo de centralizar o máximo possível de sinais e variações para cada palavra em Libras de origem regional/local. Isso permite uma grande escalabilidade, pois os usuários podem sugerir novos sinais que não estejam presentes no aplicativo acompanhando a dinâmica da Libras. Além disso, as categorias foram cuidadosamente planejadas para atender às necessidades da comunidade surda, incluindo informações que vão além das limitações iniciais, como as expressões linguísticas, que desempenham um papel crucial na cultura local, conferindo cor e vivacidade à língua[Scapolan and Silva 2018].

4. Metodologia Deste Trabalho

Esta pesquisa consiste na avaliação da usabilidade e aceitação do aplicativo Pará-LIBRAS por parte do público-alvo, para o qual planejou-se um trabalho em três etapas, as quais são: *i*) definição dos testes de usabilidade a serem aplicados, apresentados na Subseção 4.1; *ii*) seleção dos participantes para aplicação dos testes, os quais são apresentados na Subseção 4.2; e , *iii*) coleta e análise dos resultados, sobre a qual discorre-se na Subseção 4.3.

4.1. Procedimentos

Definiu-se, a partir de um levantamento na revisão de literatura, pela a aplicação de dois testes de usabilidade distintos: Teste 1, com a abordagem GQM [Rombach et al. 1994][Hussain and Kutar 2009], e Teste 2, que consiste no SUS [Brooke et al. 1996][Lourenço et al. 2022][Vlachogianni and Tselios 2022]. O Teste 1 possui como metas os seguintes objetivos (*Goals*): *i*) tornar simples usar o aplicativo, *ii*) garantir a precisão e a confiabilidade das informações no aplicativo, *iii*) minimizar o tempo necessário para a execução de tarefas no aplicativo, *iv*) prover uma interface de

usuário rica em funcionalidades necessárias, *v*) assegurar a segurança e o conforto dos usuários durante o uso do aplicativo e, *vi*) criar uma interface atraente e visualmente agradável para os usuários. Cada um desses objetivos refere-se a um eixo que será avaliado, os quais constam na primeira coluna da Tabela 1. Para cada eixo foram definidas algumas questões (*Questions*), na terceira coluna da Tabela 1, e as métricas (*Metrics*) relacionadas que ajudam a identificar se um determinado objetivo foi atingido. Assim, o teste GQM utilizado possui 18 questões afirmativas (QA)s, sendo cada uma identificada por meio de um código que descreve o eixo do qual faz parte e sua ordem dentro do eixo (segunda coluna da tabela).

Tabela 1. Eixos, questões e métricas utilizadas para o Teste 1 (GQM).

Eixos	Código	Questão	Métricas
Simplicidade	E1Q1	Eu achei fácil inserir os dados na caixa de pesquisa.	Número de respostas 4/5. Número de tentativas.
	E1Q2	No menu "Enviar Sugestão" é fácil adicionar uma nova palavra ao glossário.	
	E1Q3	No menu "Editar Categoria" é fácil adicionar uma nova categoria de palavras ao aplicativo.	
	E1Q4	O aplicativo é fácil de aprender.	
Precisão	E2Q1	O aplicativo possui informações corretas e precisas.	Quantidade de erros identificados pelos usuários.
	E2Q2	As tarefas executadas dentro do aplicativo são finalizadas corretamente na primeira tentativa.	
	E2Q3	O aplicativo funciona sem erros ou falhas.	
Tempo Gasto	E3Q1	As tarefas são executadas em poucas etapas.	Tempo médio para execução das tarefas.
	E3Q2	O aplicativo possui um bom tempo de resposta ao executar ações.	
	E3Q3	O aplicativo me economiza tempo (é prático de usar).	
Características	E4Q1	O aplicativo fornece um botão de menu (três linhas no canto superior) apropriado para a tela sensível ao toque.	Número de características rejeitadas.
	E4Q2	O aplicativo fornece todos os recursos que eu preciso.	
	E4Q3	As telas do aplicativo são divididas em seções de forma lógica/sequencial, facilitando a navegação.	
	E4Q4	O aplicativo fornece informações claras e concisas sobre seus recursos no Menu lateral.	
Segurança	E5Q1	O aplicativo não possui efeitos colaterais como cansaço visual e fadiga ou esforço muscular durante seu uso.	Quantidade de relatos de desconforto ou insegurança.
	E5Q2	Me sinto seguro ao usar o aplicativo em geral.	
Atratividade	E6Q1	A interface do aplicativo é atraente e visualmente agradável.	Quantidade de *feedback* positivo
	E6Q2	As cores e fontes usadas no aplicativo são agradáveis aos olhos.	

Cada QA pode ser respondida dentro do seguinte intervalo: 1 (Discordo Totalmente), 2 (Discordo),3 (Neutro), 4 (Concordo) e 5 (Concordo Totalmente) da escala *Likert*. Por meio deste tipo de escala de respostas é possível realizar tanto análises estatísticas descritivas quantitativas quanto qualitativas, podendo proporcionar análise de correlação [Antonialli et al. 2016].

Para a preparação do Teste 2 realizou-se *i*) a adaptação do questionário original para a língua portuguesa, conforme feito em [Lourenço et al. 2022], e em seguida, para melhor aderência ao público-alvo da avaliação de usabilidade, *ii*) realizou-se uma segunda rodada de adaptação por meio do método Comunica Simples [Sanches et al. 2022]. Esse método é definido como um conjunto de sete diretrizes (empatia, hierarquia das informações, palavra conhecida, palavra concreta, frase curta, ordem direta e disgnóstico) voltadas para a elaboração e reescrita de textos mais simples, com o objetivo de tornar acessível a compreensão do conteúdo ao público de qualquer natureza. Por exemplo, o termo "sistema (*system*)" foi substituído por "aplicativo" e "desnecessariamente complexo (*unnecessarily complex*)" por "exagerado na complexidade". A Tabela 2 mostra a versão original do teste SUS e a versão final, utilizada neste trabalho, na terceira coluna. Ambos os testes foram aplicados como em [do Nascimento 2019].

4.2. Participantes

Os participantes da pesquisa são 14 pessoas, dentre as quais, professores do ensino médio técnico federal, do ensino superior federal dos cursos de Graduação em Pedagogia, dis-

Tabela 2. Versão final do Teste 2 (SUS)

Código	Instrumento Original	Versão Final
SUSO1	*I think that I would like to use this system frequently.*	Eu acho que gostaria de usar esse aplicativo com frequência.
SUSO2	*I found the system unnecessarily complex.*	Eu acho que o aplicativo exagerado na complexidade.
SUSO3	*I thought the system was easy to use.*	Eu achei o aplicativo fácil de usar.
SUSO4	*I think that I would need the support of a technical person to be able to use this system.*	Eu acho que precisaria de ajuda de uma pessoa com conhecimentos técnicos para usar o aplicativo.
SUSO5	*I found the various functions in this system were well integrated.*	Eu acho que as várias funções do aplicativo estão muito bem integradas.
SUSO6	*I thought there was too much inconsistency in this system.*	Eu acho que o sistema apresenta muita inconsistência.
SUSO7	*I would imagine that most people would learn to use this system very quickly.*	Eu imagino que as pessoas aprenderão a usar esse aplicativo rapidamente.
SUSO8	*I found the system very cumbersome to use.*	Eu achei esse aplicativo atrapalhado de usar.
SUSO9	*I felt very confident using the system.*	Eu me senti confiante ao usar o aplicativo.
SUS10	*I needed to learn a lot of things before I could get going with this system.*	Eu precisei aprender várias coisas novas antes de conseguir usar o aplicativo.

centes de graduação dos referidos cursos, professores de Letras/LIBRAS e profissionais intérpretes de LIBRAS de uma casa legislativa municipal da cidade de Marabá, apresentados na Figura 1. O estudo realizado é de natureza quasi-experimental. Um estudo quasi-experimental é um tipo de pesquisa que compartilha algumas características tanto com estudos experimentais quanto com estudos observacionais, mas não atende completamente aos critérios de um experimento controlado [Campbell and Stanley 2015]. Este tipo de estudo foi escolhido devido a conveniência na seleção dos participantes, os quais atuavam nas mesmas instituições e utilizam LIBRAS no seu trabalho.

Os testes foram aplicados no período de cinco dias, tendo sido realizada uma seção de testes em cada uma das instituições participantes. Durante a seção de testes, inicialmente apresentava-se o protótipo de alta fidelidade em um *tablet Galaxy Tab S6 Lite LTE*, com tela de 10.4" e sistema operacional *Android*. Cada participante foi informada detalhadamente sobre os aspectos do estudo por meio do termo de consentimento livre e esclarecido (TCLE). Cada usuário participante pôde operar o aplicativo no dispositivo por até um minuto, com o objetivo de familiarizarem-se. Cada participante recebeu separadamente, uma lista contendo três tarefas para serem executadas (como em E1Q3 na Tabela 1), tendo sido acompanhadas por, pelo menos, dois pesquisadores. Os pesquisadores realizaram, por meio de observação sem intervenção, a cronometragem de tempo de execução das tarefas, registros dos números de tentativas e identificação de dificuldades no manuseio. A seção individual com cada participante durou entre 15 e 20 minutos e foi gravada em áudio. Ao final das seções de teste, os pesquisadores indagaram os participantes sobre suas impressões gerais a respeito do aplicativo, como forma de obter informações adicionais/sugestões não capturadas pelos testes. Os principais tipos de ameaças à validade da pesquisa foram identificados e tratados, conforme destacado na Tabela 3.

4.3. Coleta de Dados

Os Testes 1 e 2 foram aplicados individualmente com acompanhamento de, pelo menos, duas pessoas pesquisadoras (co-autores deste trabalho) e os participantes puderam escolher respondê-los na versão digital como um formulário do *Google Forms*, ou em versão impressa. Os dados coletados foram armazenados em uma planilha *csv*[3].

Prosseguiu-se com a fase de pré-processamento dos dados, sendo que os dados ob-

[3]Planilha disponível em `https://abrir.link/vdmAu`.

Tabela 3. Tipos de Ameaças, Detalhes e Tratamentos

Tipo de Ameaça	Detalhes da Ameaça	Tratamento da Ameaça
Instrumentação	Diferenças na forma de aplicação dos testes ou na coleta de dados.	Usar procedimentos padronizados e treinamentos para moderadores e observadores da pesquisa.
Expectativa do Pesquisador	Influência das expectativas do pesquisador sobre os resultados do teste.	Os moderadores e observadores não puderam comunicar-se ou interferir-se durante a realização das tarefas.
Seleção de Participantes	Seleção não representativa dos usuários pode limitar a generalização dos resultados.	Selecionar uma amostra diversificada (proveniente de diferentes instituições) que represente o público-alvo do aplicativo.
Configuração do Ambiente	Ambiente de teste artificial que pode não refletir o uso real do aplicativo.	Realização dos testes em ambientes que simulem o uso real, manuseando um protótipo de alta fidelidade do aplicativo.
Mono-Operacionalização	Uso de uma única medida para avaliar um constructo complexo.	Uso de múltiplas métricas e métodos (GQM e SUS) para avaliar a usabilidade de maneira abrangente.
Tamanho da Amostra	Amostra pequena pode limitar a robustez das conclusões estatísticas.	Justificar o tamanho da amostra com base em técnicas de análise qualitativa e, se possível, complementar com testes adicionais, tais como a execução de tarefas específicas.

tidos nos teste em versão impressa foram transcritos para o formulário digital. Realizou-se, ao final, a limpeza dos dados e a análise com estatística descritiva. Para analisar a confiabilidade do instrumento de coleta utilizado, em relação às respostas dos participantes, utilizou-se a variância média extraída (VME), a medida de confiabilidade composta (CC) [Valentini and Damasio 2016], e Índice de Validade de Conteúdo (IVC) para análise da validade do questionário elaborado de acordo com as diretrizes GQM. Enquanto o VME e o CC são medidas estatísticas, o IVC é uma medida que tem como base o julgamentos dos participantes. O Alfa de *Cronbach* (α), Coeficiente Ômega (Ω) e Coeficiente de Consistência de Raju (β) [Raju 1977] também foram calculados para análise da confiabilidade e consistência interna das questões [Cígler and Chvojka 2022].

São considerados bons resultados valores de α, Ω e β superiores a 0.7 [Warrens 2015]. O VME e CC são utilizados para garantir que os itens dentro de cada eixo são válidos e estão medindo o mesmo construto de maneira consistente. Já o IVC avalia a representatividade do conteúdo do questionário, garantindo que todas as áreas relevantes de usabilidade são cobertas. α é mais comumente utilizada e, por isso, é um instrumento mais fácil para comparação ajudando a identificar e melhorar itens problemáticos, mas que é influenciado pelo número de itens. Por isso considera-se ainda Ω e β, especialmente devido as variadas quantidades de questões por eixo (variabilidade nas cargas fatoriais das questões) no Teste 1 (GQM), como forma de proporcionar uma análise mais robusta da confiabilidade dos itens.

Adicionalmente, foi utilizado o coeficiente de correlação de *Pearson* [Cavalcanti et al. 2017] para identificar as correlações significativas e positivas entre as questões dos vários eixos. Quanto mais próximo de 1 for o coeficiente de *Pearson*, maior a correlação existente, sendo que é possível interpretar o resultado dentro das seguintes faixas: *Pearson* $>= 0.5$ grande correlação; *Pearson* > 0.3 correlação média; *Pearson* > 0.1 pequena correlação e *Pearson* < 0.1 significa que não há correlação. A análise final dos dados foi feita usando a ferramenta JASP e linguagem de programação *Python*, bem como a ferramenta *SUS Analisys Toolkit* [Blattgerste et al. 2022], exclusivamente para o Teste 2, nas quais os resultados foram tratados e sumarizados em tabelas e gráficos.

5. Análise dos Resultados do Estudo

A primeira análise realizada foi sobre confiabilidade e validade do Teste 1 (GQM), em relação às repostas dos participantes, visto que os itens deste tipo de teste podem variar

muito de acordo com o escopo da pesquisa. A VME e a CC para todos os eixos foi igual a 1, indicando alta consistência interna e alta qualidade do construto. Já os valores de IVC, α, Ω e β apresentaram algumas variações. No gráfico da Figura 3, cada barra azul representa o IVC de um item específico do GQM (associado ao eixo vertical da esquerda), enquanto as barras verdes representam os valores das outras métricas de confiabilidade α, Ω e β (associado ao eixo vertical da direita). Os valores de IVC variam de 0.71 a 1.00 para as diferentes questões. Quanto mais próximo de 1, maior a relevância da questão perante os participantes avaliadores do aplicativo, isto é, maior a proporção de convergência positiva de respostas (respostas 4 e 5). A questão E4Q2, que verifica se o aplicativo fornece todos os recursos de que o usuário precisa, com IVC de 0.71 sugere uma menor concordância entre os avaliadores, visto que respostas 2 e 3 foram registradas e contribuíram para manter esse IVC menor do que os das demais questões. Destaca-se o terceiro eixo, sobre tempo gasto, que atingiu um IVC de 1 em todas as suas questões (E3Q1, E3Q2, E3Q3), demonstrando que nenhum participante registrou respostas negativas ou neutras nestes itens. Os resultados são semelhantes aos encontrados em outro estudo de usabilidade com o GQM [Sim et al. 2022].

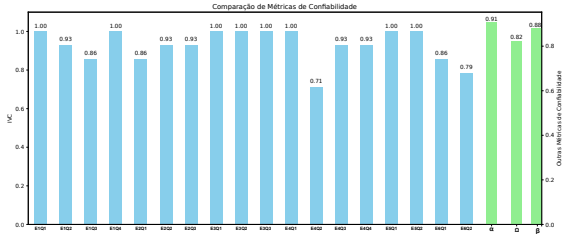

Figura 3. Confiabilidade e validade do Teste 1 com GQM

Ainda na Figura 3 observa-se que o excelente valor de α [Warrens 2015], de 0.91, indica uma alta consistência interna entre as questões do Teste 2. Há uma relação positiva entre α e a AVE: quanto mais alta a consistência interna entre os itens (refletida em um maior α), maior será a AVE, e vice-versa. Quanto maior a quantidade de itens a serem analisados, maior tende a ser o α. O resultado de α é confirmado pelo valor de $\Omega = 0.82$, significando que 82% das questão que foram propostas apresentam alta consistência interna em termos de convergências de respostas 4 e 5. Similar ao Ω e ao α, $\beta = 0.88$ confirma mais uma vez a validade por ser mais apropriado em certos contextos onde se espera variabilidade nas cargas fatoriais (por exemplo, há eixos com duas e com quatro questões). O Testes 2 (SUS) já está validado na literatura [Blattgerste et al. 2022].

A matriz de correlação de *Pearson* (Figura 4) para o GQM também é apresentada, assim feito em [Sim et al. 2022]. O objetivo é mostrar que as questões com maior correlação, avaliam aspectos semelhantes. Observa-se que muitas questões/itens do GQM apresentam uma correlação forte positiva (próxima de 1), nas cores em vermelho, indicando que quando o valor de correlação de uma dessas questões aumenta, a outra tende a aumentar também. O eixo da simplicidade demonstra que suas questões são mais fortemente correlacionadas, enquanto E4Q2 com E6Q2 têm uma correlação moderada (entre 0.3 e 0.7), o que é esperado, visto que essas questões referem-se a eixos diferentes. A presença de correlações fortes dentro de eixos (*Goals*) específicos sugere que as questões agrupadas corretamente capturam dimensões específicas do que está sendo avaliado.

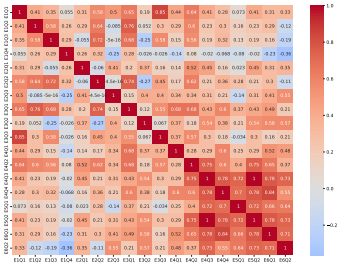

Figura 4. Matriz de Correlação de *Pearson* para o Teste 1 (GQM)

Uma segunda análise foi realizada a respeito das percepções dos participantes sobre o quanto apreciam as características avaliadas para o aplicativo. O gráfico da Figura 5A mostra o resultado por eixo ou objetivo avaliado. Quanto mais próximo do valor 5 (concordo fortemente), mais positiva é a percepção do participante em relação ao eixo avaliado. Sobre a Simplicidade do aplicativo, há uma forte consistência com a maioria das respostas positivas, sendo que a questão E1Q3, que verifica se é fácil adicionar uma nova categoria de palavras a partir do acesso a opção ao menu "Editar Categoria", apresentou maior variabilidade. A questão E1Q4 demonstra menor variabilidade, o que significa que os participantes acharam o aplicativo fácil de aprender. A Precisão foi bem avaliada, sendo que houve uma maior variabilidade nas respostas para a questão que verifica se as informações dos símbolos e representação em LIBRAs estão corretas e precisas. Alguns participantes identificaram erros de semântica em algumas expressões presentes no glossário. Entretanto, o funcionamento do aplicativo ocorre sem erros ou falhas, conforme mostra o resultado para E2Q3.

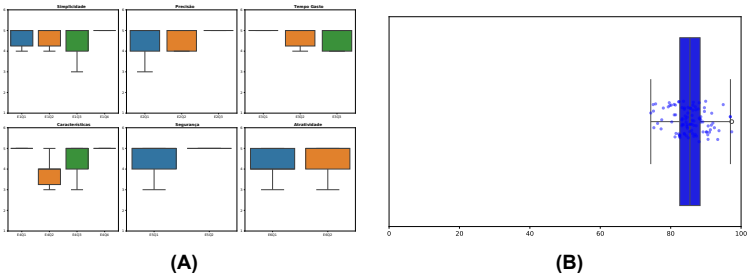

Figura 5. Resultados dos testes (A) GQM e (B) SUS

Ainda no gráfico da Figura 5A, os participantes também tiveram boa percepção sobre o Tempo Gasto para utilizar as funções principais do aplicativo e executar algumas tarefas. Sobre as Características, foi perguntado na questão E4Q2 se o aplicativo fornece todos os recursos necessários para o usuário, e a maioria das respostas se concentrou nos valores 3 e 4. Esse resultado reflete que os participantes sentiram falta de uma apresentação das expressões também em formato de vídeo ou *gifs*. A Segurança e a Atratividade do aplicativo também foram bem avaliadas, mas com alguns *outliers* indicando respostas menores. Esses resultados sugerem que uma provável mudança no esquema de cores e disposição das imagens poderia agradar ainda mais os participantes.

Já a Figura 5B mostra o resultado do Teste 2 (SUS). A pontuação (*score*) média

foi de 93, indicando uma excelente usabilidade do aplicativo, com desvio padrão de 7.3, indicando consistência na percepção da usabilidade entre os participantes devido a baixa variabilidade. Assim, o aplicativo também recebeu uma pontuação muito alta no SUS, indicando que os participantes consideraram o aplicativo altamente usável e eficiente para o propósito. A consistência das respostas reforça a confiabilidade dos resultados obtidos.

Após apresentada a análise dos eixos (*Goals* do GQM) e das questões/itens (*Questions* do GQM), discorre-se a seguir sobre as métricas avaliadas com o teste (*Metrics* do GQM), que foram destacadas na terceira coluna da Tabela 1. A Tabela 4 mostra, em relação à Simplicidade, que 1 em cada 4 participantes precisou executar alguma tarefa mais de uma vez para uma execução bem sucedida. Foram identificados, em média, dois erros, os quais se referem a símbolos em Libras com significados errados para determinadas expressões, como foi o caso das representações para as letras "M"e "N". O tempo médio de execução das tarefas foi de 33 segundos, porém, as tarefas que requerem digitação e busca de imagens na galeria do dispositivo tendem a levar uma tempo bem maior. Não foram ouvidos relatos de incômodos com as características pre-existentes no aplicativo, tampouco de questões de insegurança e desconforto. No geral os *feedbacks* foram positivos, com cerca de 90% de respostas 4 e 5. Resultados de *feedbacks* como este também foram obtidos com o mesmo método GQM em [McGinn et al. 2018] considerando outros contextos de aplicação.

Tabela 4. Métricas computadas no teste de usabilidade do aplicativo

Eixo	**Métrica**	**Média**
Simplicidade	Número de Tentativas	0,24
Precisão	Quantidade de Erros Identificados pelo Usuário	2
Tempo Gasto	Tempo Médio a Executar Tarefas	33
Características	Características Rejeitadas	0
Segurança	Relatos de Desconforto	0
Atratividade	*Feedback* Positivo	0,90

Alguns dos objetivos chave que foram satisfatoriamente atingidos nesta pesquisa incluem a garantia de simplicidade, minimização do tempo necessário para execução de tarefas, garantia de conforto e a segurança dos usuários (destacados na Subseção 4.1), os quais se relacionam com os eixos simplicidade, tempo gasto, características e segurança. Porém, como algumas falhas foram identificadas sobre a precisão das informações veiculadas no aplicativo, e devido as sugestões de funcionalidades adicionais sugeridas pelos participantes dos testes, os eixos sobre precisão e atratividade precisam ser melhor explorados. Uma possibilidade de solução é reelaborar as questões/itens desses eixos para aumentar a consistência interna e incluir novas questões para capturar com mais detalhes as impressões dos usuários participantes. Se essa hipótese for válida, então as consequências serão valores mais altos de IVC e outras medidas estatísticas, bem como obtenção de um *score* bem mais elevado na escala SUS.

6. Considerações Finais

Este trabalho apresentou uma avaliação de usabilidade de um aplicativo móvel proposto, que oferece um glossário bilíngue português-LIBRAS, e permite que os usuários possam adicionar novas expressões com seus significados. A avaliação consistiu em dois instrumentos: GQM e SUS. Apresentou-se uma análise da validade e confiabilidade dos

resultados do teste GQM considerando variância média extraída (VME), a medida de confiabilidade composta (CC), Indice de Validade de Conteúdo (IVC), Alfa de *Cronbach*, Coeficiente Ômega e Coeficiente de Consistência de Raju.

Os principais resultados do GQM foram a grande aceitação das características do aplicativo e impressões positivas acerca da simplicidade, segurança e atratividade do protótipo avaliado. Quanto ao SUS, o aplicativo obteve grande quantidade de avaliações positivas, alcançando uma pontuação média de 93 dentro da sua escala de pontuações.

As principais limitações do trabalho são as seguintes: *i*) embora tenha sido feito um esforço para reescrever o Teste 2 (SUS) em linguagem simples, seguindo metodologias específicas nesse contexto, a compreensão das questões por parte dos respondentes não pôde ser validada; *ii*) a dificuldade em atingir um público-alvo maior se deu devido a baixa aderência de pessoas que se enquadrassem no perfil da pesquisa, presentes nos órgãos públicos que atuaram como parceiros, e *iii*) considerando-se o tamanho da amostra (14 participantes), destaca-se que os achados evidenciam apenas o contexto em que o estudo foi realizado, não sendo possível generalizações nesse sentido.

Como trabalhos futuros espera-se adicionar as melhorias sugeridas pelos participantes dos testes, tais como utilização de vídeos e *gifs*. Espera-se ainda o desenvolvimento de novas funcionalidades, incluindo audiodescrição, bem como expandir a avaliação para um público-alvo maior e mais diversificado.

Referências

Antonialli, F., Antonialli, L. M., and Antonialli, R. (2016). Usos e abusos da escala likert: estudo bibliométrico nos anais do enanpad de 2010 a 2015. In *Congresso de Administração, Sociedade e Inovação*, volume 1, pages 12–02, Volta Redonda, RJ. Universidade Federal Fluminense.

Blattgerste, J., Behrends, J., and Pfeiffer, T. (2022). A web-based analysis toolkit for the system usability scale. In *Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '22, page 237–246, New York, NY, USA. Association for Computing Machinery.

Brooke, J. et al. (1996). Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7.

Campbell, D. T. and Stanley, J. C. (2015). *Experimental and quasi-experimental designs for research*. Ravenio books, Cambridge, England.

Cavalcanti, A., Ferreira, R., Dionísio, M., Neto, S., Passero, G., and Miranda, P. (2017). Uma nova abordagem para detecçao de plágio em ambientes educacionais. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educaçao-SBIE)*, volume 28, page 1177.

Cígler, H. and Chvojka, E. (2022). Reliability estimation in tests composed of two items only: Admissible and plausible reliability ranges. *PsyArXiv. February*, 16.

Câmara Municipal de Marabá (CMM) (2024). Câmara lança glossário de libras para apoiar comunidade surda. Disponível em `https://maraba.pa.leg.br/institucional/noticias/`

camara-lanca-glossario-de-libras-para-apoiar-comunidade-surda. Acesso em 10 de junho de 2024.

do Nascimento, H. R. (2019). *Investigação de teste de usabilidade para aplicações móveis*. PhD thesis, Universidade Estadual de Campinas.

Friedrich, M. A. (2019). Glossário em libras: uma proposta de terminologia pedagógica (português-libras) no curso de administração da ufpel. Master's thesis, Universidade Federal de Pelotas.

Galvão, L., García, L., and Felipe, T. (2020). Concepção de jogos educativos para crianças surdas baseados na educação infantil bilíngue: Um estudo de caso de avaliação da metodologia cajedus. In *Anais do XXXI Simpósio Brasileiro de Informática na Educação*, pages 592–601, Porto Alegre, RS, Brasil. SBC.

Gesser, A. (2012). O ouvinte e a surdez: sobre ensinar e aprender a libras. *São Paulo: Parábola Editorial*.

Haôr, Victor (2023). Caes: Estudantes recebem atendimento especializado na área da surdez na sede do centro e nas escolas públicas. Disponível em https://maraba.pa.gov.br/caes-atendimento-especializado-surdez/. Acesso em 10 de junho de 2024.

Hussain, A. and Kutar, M. (2009). Usability metric framework for mobile phone application. *PGNet, ISBN*, 2099:978–1.

Kaya, A., Ozturk, R., and Altin Gumussoy, C. (2019). Usability measurement of mobile applications with system usability scale (sus). In *Industrial Engineering in the Big Data Era: Selected Papers from the Global Joint Conference on Industrial Engineering and Its Application Areas, GJCIE 2018, June 21–22, 2018, Nevsehir, Turkey*, pages 389–400. Springer.

Lourenço, D. F., Carmona, E. V., and Lopes, M. H. B. d. M. (2022). Translation and cross-cultural adaptation of the system usability scale to brazilian portuguese. *Aquichan*, 22(2).

McGinn, C., Bourke, E., O'Kelly, T., and Cullinan, M. F. (2018). Adapting the goals/questions/metrics (gqm) method for applications in robot design. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4746–4751.

Oliveira, A. M., Cabral, A., Barreto, G., Nascimento, J., Rodrigues, S., and Viana, F. (2020). Software educativo contakg bilíngue. In *Anais dos Workshops do X Congresso Brasileiro de Informática na Educação*, pages 221–226, Porto Alegre, RS, Brasil. SBC.

Ouedraogo, E. O. G., Rodrigues, E. G., and Ouedraogo, A. (2020). Glossário jurídico em libras: sinal, discurso e linguagem cinematográfica. *Línguas e Instrumentos Línguísticos*, 23(46):200–223.

Pereira, C. S. (2021). Para um glossário bilíngue (português-libras) de ortodontia.

Raju, N. S. (1977). A generalization of coefficient alpha. *Psychometrika*, 42(4):549–565.

Rombach, D., Basili, V., and Caldiera, G. (1994). The goal question metric approach. *Encyclopedia of Software Engineering*, 2:528–532.

Sanches, E. C. P., Bueno, J., et al. (2022). Uso da linguagem simples como prática no design da informação e design inclusivo. *Coletânea de estudos do PPGDesign/UFPR: Novos horizontes da pesquisa em design*, pages 231–245.

Scapolan, B. A. and Silva, I. A. L. (2018). Expressões indiomaticas português-libras: (in)traduzibilidade. *Libras, Lexicografia e Cultura*, 45(Especial):332–350.

Sim, Y. W., WaiShiang, C., Phang, P., Lam, K.-C., Phang, E., and binti Jali, N. (2022). Goal question metric as an interdisciplinary tool for assessing mobile learning application. *International Journal of Advanced Computer Science and Applications*, 13(5).

Stedile, C. G. S., Fleury, G. S., and de Souza Ribeiro, M. W. (2019). Poster: Desenvolvimento de um ambiente virtual imersivo para auxiliar em práticas de falar em publico. In *Anais Estendidos do XXI Simpósio de Realidade Virtual e Aumentada*, pages 43–44. SBC.

Valentini, F. and Damasio, B. F. (2016). Average variance extracted and composite reliability: Reliability coefficients/variancia media extraida e confiabilidade composta: Indicadores de precisao. *Psicologia: Teoria e Pesquisa*, 32(2):NA–NA.

Vlachogianni, P. and Tselios, N. (2022). Perceived usability evaluation of educational technology using the system usability scale (sus): A systematic review. *Journal of Research on Technology in Education*, 54(3):392–409.

Warrens, M. J. (2015). Some relationships between cronbach's alpha and the spearman-brown formula. *Journal of Classification*, 32:127–137.