

CRIAÇÃO DE UM GLOSSÁRIO BILÍNGUE LIBRAS-PORTUGUÊS:
UMA INICIATIVA PARA PROMOVER ACESSIBILIDADE COMUNICACIONAL

***Creation of a Bilingual LIBRAS-Portuguese Glossary:
An Initiative to Promote Communicational Accessibility***

DOI: 10.14393/LL63-v37S-2026-8

Ronnie Fagundes de Brito^{1*}

Gabriela Wick-Pedro^{2**}

RESUMO: Este estudo tem como objetivo descrever as etapas iniciais da elaboração de um glossário bilíngue em português do Brasil e Língua Brasileira de Sinais (Libras), projeto atualmente em desenvolvimento, visando promover o acesso à informação científica e tecnológica por parte da comunidade surda e outros grupos com barreiras linguísticas e comunicacionais. A metodologia adotada inclui Processamento de Linguagem Natural (PLN) para a extração automática dos termos, além da análise semântica e validação dos especialistas, tanto ouvintes quanto surdos. Os resultados obtidos envolvem a identificação, categorização e validação de 130 termos científicos, organizados em seis *clusters* temáticos, com estabelecimento e relações semânticas e estruturação em base de dados terminológica. Assim, este trabalho reforça o potencial das tecnologias linguísticas na difusão da inclusão e democratização do conhecimento científico.

PALAVRAS-CHAVE: Glossário bilíngue. Processamento de Linguagem Natural. Divulgação Científica. Acessibilidade comunicacional. Inclusão social.

ABSTRACT: This study aims to describe the initial stages in the development of a bilingual glossary in Brazilian Portuguese and Brazilian Sign Language (Libras), a project currently in progress, intended to promote access to scientific and technological information for the deaf community and other groups with linguistic and communicational barriers. The methodology includes Natural Language Processing (NLP) techniques for the automatic extraction of terms, as well as semantic analysis and expert validation conducted by both hearing and deaf specialists. The results obtained involve the identification, categorization, and validation of 130 scientific terms, organized into six thematic clusters, with the establishment of semantic relations and the structuring of a terminological database. Thus, this work reinforces the potential of linguistic technologies in fostering inclusion and the democratization of scientific knowledge.

KEYWORDS: Bilingual glossary. Natural Language Processing. Scientific communication. Communicational accessibility. Social inclusion.

^{1*} Doutor. Instituto Brasileiro de Informação em Ciência e Tecnologia. ORCID: 0000-0002-3979-603X. E-mail para contato: ronniebrito(AT)ibict.br.

^{2**} Doutora. Instituto Brasileiro de Informação em Ciência e Tecnologia. ORCID: 0000-0002-7332-4482. E-mail para contato: gabrielawick(AT)ibict.br.

1 Introdução

A divulgação científica desempenha um papel estratégico na circulação do conhecimento, atuando como mediadora entre o saber acadêmico e sua apropriação por diferentes públicos (Massarani; Moreira; Brito, 2002; Björk, 2005; Bueno, 2010). Mais do que popularizar a ciência, trata-se de democratizar o acesso à informação, fortalecer a alfabetização científica e promover a cidadania crítica. Contudo, persistem desafios estruturais para alcançar públicos historicamente excluídos, como pessoas com deficiência, em especial, a comunidade surda brasileira (Oliveira, 2015).

A complexidade conceitual dos termos técnicos empregados na ciência, muitas vezes ausentes de sinais codificados em Libras, agrava esse cenário, limitando a compreensão mesmo entre surdos proficientes na leitura e escrita em português. Tais dificuldades evidenciam a urgência de se desenvolver materiais terminológicos bilíngues que atendam às especificidades linguísticas e culturais da comunidade surda, promovendo uma verdadeira acessibilidade comunicacional (Prometi; Costa; Tuxi, 2015; Tuxi, 2015; Santos, 2018).

Nesse contexto, o Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict)³, atuando na democratização da informação científica, desenvolve ações para ampliar a acessibilidade informacional a grupos com vulnerabilidade linguística. Uma dessas ações é a criação de um glossário bilíngue Libras-Português, focado em termos científicos frequentes no portal do *Canal Ciência*⁴. A iniciativa busca superar barreiras linguísticas enfrentadas por pessoas surdas no acesso a conteúdos científicos, promovendo práticas inclusivas na comunicação pública da ciência. Trata-se de um projeto em andamento, cujas etapas iniciais são descritas nesta pesquisa.

O glossário foi desenvolvido com base em uma metodologia interdisciplinar, combinando técnicas de Processamento de Linguagem Natural (PLN), análise semântica e validação colaborativa com a participação de especialistas ouvintes e surdos. Essa abordagem

³ Órgão federal responsável por ações, políticas e infraestrutura voltadas à produção, organização, disseminação e acesso à informação científica e tecnológica no Brasil.

⁴ O portal *Canal Ciência* é uma plataforma de divulgação científica mantida pelo *Instituto Brasileiro de Informação em Ciência e Tecnologia* (Ibict), dedicada à disseminação de conteúdos e materiais educativos voltados à popularização da ciência no Brasil. Disponível em: <https://canalciencia.ibict.br/>. Acesso em: 20 jun. 2025.

busca garantir a acurácia conceitual e a adequação linguística das entradas, respeitando tanto os princípios da terminologia científica quanto os parâmetros culturais e visuais da Libras.

Este artigo descreve as primeiras etapas do processo metodológico de construção do glossário bilíngue e apresenta breves considerações sobre os desafios enfrentados e sobre implicações sociais e comunicacionais dessa iniciativa. Mais do que um produto linguístico, o glossário configura-se como uma ferramenta de inclusão e transformação social, que pode servir de referência para outras ações voltadas à promoção de acessibilidade linguística em distintas áreas do conhecimento científico.

2 Pressupostos teóricos

A construção de um glossário bilíngue Libras–Português baseia-se em um conjunto de fundamentos teóricos que dialogam com diferentes áreas do conhecimento, como a linguística aplicada, a tecnologia linguística e a educação inclusiva. Esses referenciais ajudam a compreender as particularidades linguísticas da comunidade surda e fornecem suporte para o desenvolvimento de instrumentos que ampliem o acesso à informação científica. Mais do que uma simples lista de termos, o glossário é pensado como um recurso que promove acessibilidade comunicacional e contribui para a inclusão social e educacional. Assim, os pressupostos teóricos que orientam este trabalho sustentam tanto os caminhos metodológicos escolhidos quanto os objetivos sociais mais amplos que envolvem a democratização do conhecimento.

2.1 Acessibilidade comunicacional e inclusão linguística

A acessibilidade comunicacional é entendida como um direito fundamental das pessoas surdas ao acesso equitativo à informação e ao conhecimento (Brasil, 2004; Strobel, 2008). Garantir esse acesso significa assegurar condições reais de compreensão e participação em contextos comunicativos, independentemente das habilidades sensoriais ou linguísticas dos indivíduos. No Brasil, a Língua Brasileira de Sinais (Libras) foi reconhecida como língua oficial pelo Estado com a promulgação da Lei n.º 10.436/2002, devendo ser garantida como meio legítimo de acesso à informação.

Apesar desse reconhecimento legal, ainda há escassez de recursos terminológicos e educacionais em Libras que permitam o acesso pleno a conteúdos científicos e técnicos. Muitos termos técnicos não possuem equivalente direto ou amplamente difundido na Libras, o que dificulta a compreensão por parte de surdos, mesmo aqueles proficientes em leitura e escrita em português. Nesse contexto, a elaboração de um glossário bilíngue surge como uma iniciativa estratégica para reduzir essas barreiras linguísticas e promover a democratização do saber científico.

2.2 Processamento de Linguagem Natural e metodologia linguística

O desenvolvimento tecnológico tem oferecido novas possibilidades para a criação de recursos linguísticos voltados à inclusão. O uso de PLN permite identificar termos técnicos frequentes em textos de divulgação científica de forma sistemática e replicável. Técnicas como análise de *n-gramas*⁵, uso de *stoplists*⁶ e *parser*⁷ morfossintático ajudam a extrair candidatos a termos relevantes, formando listas inicialmente tratadas com auxílio de *softwares*, como o *AntConc*⁸ (Anthony, 2023). Essa abordagem automatizada é essencial para lidar com grandes *corpora* textuais e garantir uma seleção eficiente de unidades lexicais que representem adequadamente o domínio técnico-científico explorado.

A metodologia adotada neste projeto teve como base os pressupostos teóricos e práticos propostos por Dias-da-Silva (2006), que apresenta uma abordagem estruturada para o desenvolvimento de recursos linguísticos computacionais. Segundo a autora, a construção de bases de dados linguísticas envolve três domínios principais: o domínio linguístico, responsável pela definição dos aspectos teóricos e descritivos da língua; o domínio

⁵ *N-gramas* são unidades formadas por sequências contíguas de *n* elementos linguísticos (palavras ou caracteres), utilizadas para a modelagem estatística e a identificação de padrões recorrentes em corpora no âmbito do Processamento de Linguagem Natural.

⁶ Os *stoplists* são conjuntos de itens lexicais de alta frequência, geralmente palavras funcionais, que são excluídos das etapas de pré-processamento textual a fim de reduzir ruído e aprimorar a relevância das informações extraídas.

⁷ O *parser* é um mecanismo computacional destinado à análise da estrutura sintática de enunciados, responsável por identificar hierarquias, dependências e relações formais entre os constituintes de uma sentença.

⁸ *AntConc* é um *software* gratuito de análise de *corpora* desenvolvido por Laurence Anthony, empregado na geração de concordâncias, listas de frequência e estatísticas lexicais para pesquisas em Linguística de *Corpus*. Disponível em: <https://www.laurenceanthony.net/software/antconc/>.

representacional, que se ocupa da forma como as informações serão organizadas e visualizadas; e o domínio implementacional, voltado para a aplicação tecnológica e a disponibilização do produto final.

Embora a metodologia em questão tenha sido originalmente concebida para o desenvolvimento de sistemas de PLN, ela foi posteriormente adaptada por Di Felippo e Almeida (2010) para a construção de *wordnets* terminológicas⁹ em português brasileiro, evidenciando a necessidade de ajustes conforme o objetivo final. No presente trabalho, essa abordagem foi direcionada à criação de um glossário bilíngue voltado à acessibilidade comunicacional, priorizando-se as etapas relacionadas ao domínio linguístico e representacional. Tal enfoque assegurou consistência conceitual, utilidade pedagógica e alinhamento com as demandas dos públicos surdo e ouvinte.

Apesar de o PLN oferecer suporte metodológico relevante, seu uso exige complementação com análises qualitativas e validações colaborativas, uma vez que a tradução de termos técnicos nem sempre apresenta equivalência direta entre línguas, exigindo adaptações ou a criação de novos sinais. Além disso, o PLN pode contribuir para o mapeamento de relações semânticas, possibilitando a organização dos termos em redes conceituais dinâmicas e acessíveis. Isso torna possível não apenas a estruturação do glossário de forma clara, mas também sua expansão contínua, à medida que novos conceitos e sinais forem integrados ao repertório da comunidade surda.

3 Metodologia

A construção de um glossário bilíngue português-Libras voltado à divulgação científica envolve uma metodologia estruturada e multidisciplinar, integrando técnicas de PLN, análise semântica e validação colaborativa por especialistas ouvintes e surdos. A abordagem adotada é fundamentada na metodologia proposta por Dias-da-Silva (2006), composta por três domínios principais: linguístico, representacional e implementacional. Embora originalmente desenvolvida para sistemas de PLN, essa abordagem foi ajustada conforme as necessidades

⁹ Redes léxico-conceituais estruturadas para representar termos especializados e suas relações semânticas, organizadas em modelos hierárquicos e associativos que apoiam aplicações em Terminologia e Processamento de Linguagem Natural.

do produto final. Aqui, um recurso terminológico bilíngue destinado à promoção da acessibilidade comunicacional.

O *corpus* utilizado foi extraído automaticamente do portal *Canal Ciência*, serviço de divulgação científica vinculado ao Ibict. Totalizando 323.707 *tokens*¹⁰, constitui uma base de dimensão adequada aos objetivos desta pesquisa terminológica, possibilitando a identificação de termos técnicos frequentes nas áreas de ciência, tecnologia, saúde, meio ambiente, geografia e diversidade populacional.

3.1 Limpeza e preparação do corpus

A primeira etapa do processo envolve a limpeza e preparação do *corpus*. Essa etapa desempenha um papel essencial na construção de um *corpus* limpo e consistente, adequado para as análises subsequentes. Durante essa fase, são realizadas várias ações para garantir que os textos utilizados sejam de alta qualidade e livres de elementos que possam interferir na extração de informações relevantes. Essa fase incluiu:

- 1) **Remoção de dados irrelevantes e ruídos:** foram removidos do *corpus* elementos textuais sem valor terminológico, como marcadores de formatação e referências externas (“Figura:”, “Tabela:”, “Fonte:”, “Crédito:”, “Foto”), por representarem ruídos que poderiam comprometer a análise lexical.
- 2) **Exclusão de links e endereços eletrônicos:** URLs, links externos e endereços de e-mail foram eliminados por não contribuírem para a análise terminológica e por poderem introduzir distorções no processamento automático do *corpus*.
- 3) **Remoção de enumerações e pontuações irrelevantes:** foram excluídas enumerações e símbolos como “1)”, “a)”, “–”, “:”, “•”, por não acrescentarem valor semântico ao *corpus* e poderem interferir negativamente na extração terminológica.

¹⁰ *Tokens* são unidades mínimas resultantes do processo de tokenização, correspondentes às ocorrências individuais de palavras ou símbolos em um *corpus*, empregadas como base para análises estatísticas e linguísticas no Processamento de Linguagem Natural.

- 4) **Eliminação de tags HTML:** tags de formatação HTML, como , e <i>, foram removidas para assegurar que o *corpus* contenha apenas texto bruto, favorecendo uma análise terminológica mais precisa.
- 5) **Correção ortográfica e normalização:** erros de grafia foram corrigidos e variações de escrita padronizadas (por exemplo, “acucar”, “açucar” e “açúcar”) para evitar inconsistências na extração e categorização dos termos.
- 6) **Remoção de espaços em branco:** espaços em excesso entre palavras ou no final de frases foram eliminados para garantir a uniformidade do texto e evitar distorções no processamento.

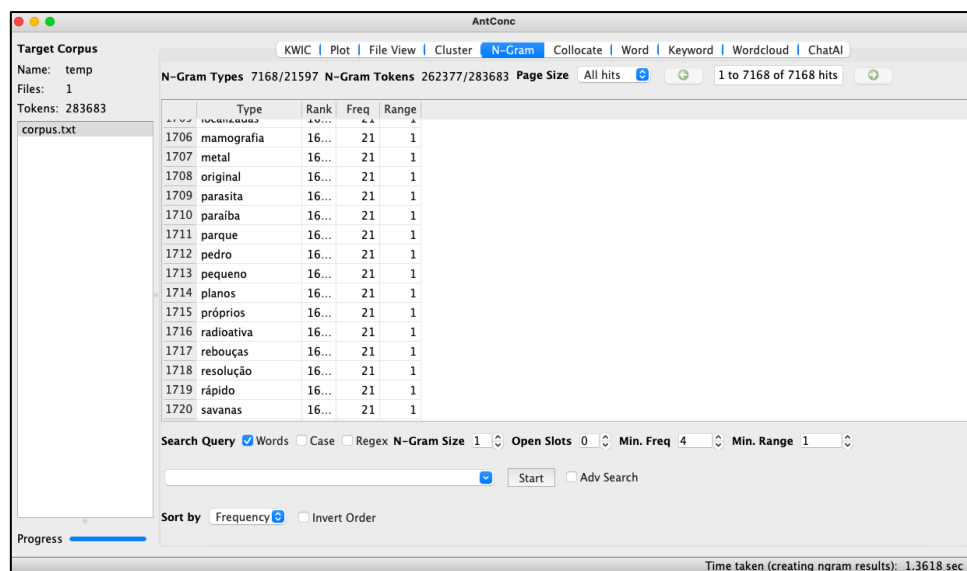
Essas etapas de limpeza, embora demandem tempo, são essenciais para assegurar a qualidade do *corpus* e a precisão das análises, impactando diretamente a eficácia da extração terminológica e a consistência do glossário final.

3.2 Extração automática de candidatos a termos

Na segunda fase, a extração dos candidatos a termos foi realizada automaticamente por meio de técnicas de PLN. Para isso, foi utilizada a ferramenta AntConc (Anthony, 2023) para identificar combinações frequentes de palavras no corpus, incluindo tanto unigramas (palavras isoladas) quanto n-gramas com até oito palavras. A extração de unigramas justifica-se pela natureza do domínio científico analisado: termos técnicos simples, como *parasita*, *radioativa* ou *resolução*, ocorrem com frequência relevante nos textos de divulgação do *Canal Ciência* e compõem parte significativa do vocabulário especializado. Todos os itens extraídos foram tratados como candidatos a termos, sujeitos a filtragem e validação posteriores (seções 3.3 e 3.4), uma vez que nem todo item frequente possui valor terminológico para o domínio. Os critérios adotados foram a frequência mínima de ocorrência, calculada conforme Lopes et al. (2009), e a pertinência contextual ao campo da divulgação científica. Ao final da extração, os candidatos foram organizados em dois grupos: 1-gramas (palavras isoladas) e n-gramas com até cinco palavras, embora o AntConc (Anthony, 2023) estivesse configurado para identificar estruturas de até oito palavras, sequências mais longas mostraram-se pouco representativas após a filtragem. A Figura 1 ilustra, de forma parcial, a interface do AntConc

com candidatos a termos simples (unigramas) extraídos do corpus, configurados com frequência mínima de 4 ocorrências.

Figura 1 – Exemplo de extração de termos



Fonte: Candidatos a termos simples (unigramas) extraídos do corpus pelo AntConc, com frequência mínima de 4 ocorrências (exibição parcial).

Na segunda fase, a extração dos candidatos a termos foi realizada automaticamente por meio de técnicas de PLN, com o uso da ferramenta AntConc (Anthony, 2023). Foram extraídos n-gramas de até oito palavras, capturando tanto itens simples quanto compostos. Em seguida, as listas passaram por limpeza e refinamento para assegurar a relevância (e a precisão) dos candidatos selecionados, conforme descrito na seção 3.3.

3.3 Limpeza e refinamento das listas de candidatos a termos

Na terceira etapa da extração de termos, foi realizado um processo de refinamento e filtragem das listas de n-gramas, com o objetivo de selecionar apenas os candidatos mais relevantes ao glossário. Inicialmente, aplicou-se um corte de frequência baseado na fórmula $\text{número de palavras do corpus} / 100.000 + 1$, proposta por Lopes et al. (2009). Aplicada ao corpus deste estudo (323.707 tokens), a fórmula resulta em 4,24, adotando-se, portanto, o valor mínimo de 4 ocorrências. Essa filtragem eliminou n-gramas com baixa frequência e foi

complementada por diversas iterações de limpeza. Uma *stoplist*¹¹ foi utilizada para excluir palavras de uso geral (como preposições, artigos e pronomes), que não possuem valor técnico no domínio científico analisado.

Além disso, utilizou-se um parser para identificar as categorias morfosintáticas dos n-gramas extraídos, com base na anotação automática realizada pelo *Mac-Morpho* do *NLTK*¹². Esse procedimento permitiu a remoção de termos inadequados ao glossário, como adjetivos estrangeiros, advérbios, conjunções, números e pronomes. Também foram eliminadas sequências iniciadas ou terminadas por conectivos (como “*da universidade*”), nomes próprios isolados, expressões genéricas ou irrelevantes e termos sem contexto claro.

3.4 Validação dos termos pelos especialistas

A validação dos termos extraídos foi essencial para garantir a precisão conceitual e a relevância do glossário. O processo ocorreu em duas fases, ambas com índice mínimo de concordância de 80%. Na primeira, cinco especialistas ouvintes analisaram 1.284 candidatos a termos, dos quais 98 foram validados. Na segunda, a equipe foi ampliada para oito especialistas (seis ouvintes e dois surdos), com peso dobrado atribuído às avaliações dos surdos para assegurar a representatividade da comunidade usuária da Libras; dessa rodada, 130 termos foram validados

Esse processo criterioso assegurou a adequação técnica e científica dos termos, ao mesmo tempo em que considerou a acessibilidade linguística e cultural do público-alvo. Com base no *feedback* dos especialistas, ajustes e refinamentos foram incorporados ao glossário, elevando sua qualidade final. A duração dessa etapa foi determinada pela disponibilidade dos avaliadores e pela natureza colaborativa do trabalho.

3.5 Edição do mapa conceitual e categorização dos termos

¹¹ *Stoplist* é uma lista de palavras gramaticais ou de alta frequência e baixo valor informacional, como preposições, artigos, pronomes e conjunções, que são excluídas do processamento automático por não contribuírem para a identificação de termos técnicos.

¹² NLTK (Natural Language Toolkit) é um conjunto de bibliotecas de software que auxiliam na análise e tratamento linguístico computacional <<https://www.nltk.org/>>.

Nessa etapa, a edição do mapa conceitual e a categorização dos termos foram essenciais para organizar e estruturar o conhecimento de forma clara e acessível. Inicialmente, os 98 termos validados pelos seis especialistas falantes foram agrupados em seis *clusters* temáticos¹³, definidos com base nas áreas de conhecimento mais relevantes para o glossário. Os *clusters* estabelecidos foram:

1. **Pesquisa e Divulgação:** abrange termos relacionados à produção científica, divulgação do conhecimento e ferramentas utilizadas nesse processo.
2. **Ambiente e Sustentabilidade:** reúne termos sobre recursos naturais, ecologia e práticas sustentáveis, incluindo temas como “desenvolvimento sustentável” e “meliponicultura”¹⁴.
3. **Ciência e Tecnologia:** foca em termos ligados à inovação tecnológica e suas aplicações científicas em diferentes áreas.
4. **Saúde e Genética:** agrupa termos relacionados à biologia, saúde pública e genética molecular, abrangendo temas essenciais para o campo da saúde.
5. **Diversidade e Populações:** agrupa termos sobre diversidade genética, social e ecológica, considerando tanto aspectos biológicos quanto sociais.
6. **Geografia e Urbanização:** inclui termos ligados à geografia física e humana, além de aspectos sociais do desenvolvimento urbano.

Após a categorização, foi estabelecida uma relação semântica entre os termos e seus respectivos *clusters*, utilizando uma estrutura de triplas (sujeito-predicado-objeto). Essas triplas são essenciais para esclarecer a interconexão entre os termos, facilitando a compreensão das relações conceituais. Exemplos de triplas incluem: "pesquisa, é divulgada por, divulgação científica" e "meio ambiente e sustentabilidade, combate, aquecimento

¹³ Referem-se a *clusters* temáticos os agrupamentos de elementos que compartilham características semânticas ou funcionais semelhantes, constituídos por procedimentos de análise quantitativa ou qualitativa, com o propósito de identificar padrões, categorias ou áreas de convergência em um corpus ou conjunto de dados. Diferentemente dos n-gramas, sequências de palavras extraídas automaticamente com base em critérios de frequência, os *clusters* são definidos pelos pesquisadores com base em afinidade conceitual entre os termos validados.

¹⁴ A meliponicultura é a prática de criação e manejo de abelhas sem ferrão (Meliponini), tradicional em diversas regiões do Brasil e frequentemente associada à conservação ambiental. No presente estudo, trata-se de um dos termos especializados extraídos e analisados no corpus.

global". Essas triplas ajudam a mapear as conexões entre os termos e como eles se relacionam em um contexto maior.

Para visualizar essas conexões, foi criada uma rede semântica que interliga os termos dentro dos *clusters*, utilizando a ferramenta *yED Graph Editor*¹⁵, que permitiu organizar e representar graficamente essas relações de forma clara e eficiente. Com essa rede estabelecida, o próximo passo será o gerenciamento da base de dados terminológicos, integrando os termos validados, suas definições e interconexões para garantir fácil acesso e atualização constante.

3.6 Gerenciamento da base de dados terminológicos

Na última etapa, o foco é o desenvolvimento e a gestão da base de dados terminológica. Esta base será alimentada com as informações coletadas nas fases anteriores, incluindo definições dos termos, exemplos de uso e traduções para Libras. O preenchimento das fichas terminológicas completará o glossário, tornando-o acessível tanto para a comunidade científica quanto para a população surda, que necessita da tradução de termos técnicos para Libras.

Cada entrada do verbete será estruturada de maneira detalhada, como exemplificado a seguir:

- **Termo em português:** registro da forma canônica do termo, acompanhado da categoria gramatical (ex.: substantivo, adjetivo) e de uma definição clara que situe o conceito no domínio da divulgação científica.
- **Termo em Libras:** representação visual do sinal correspondente, com descrição dos parâmetros fonológicos da Libras: configuração de mão, ponto de articulação, movimento e expressão facial. Quando não houver sinal consolidado, será proposto um novo sinal em colaboração com especialistas surdos.
- **Exemplo de uso:** frase ou contexto extraído do corpus ilustrando o emprego do termo em português, acompanhado de vídeo demonstrativo do sinal em Libras.

¹⁵ O *yED Graph Editor* é um *software* de visualização e modelagem gráfica utilizado para criar, organizar e editar diagramas, permitindo a representação estruturada de redes conceituais, taxonomias e relações semânticas. No presente estudo, foi empregado para a representação visual das estruturas terminológicas.

- **Notas explicativas:** informações complementares sobre variações de uso, sinônimos, termos relacionados ou especificidades culturais relevantes para a comunidade surda.
- **Fonte:** referência bibliográfica ou documental utilizada para embasar a definição adotada.
- **Domínio ou área temática:** indicação do *cluster* temático ao qual o termo pertence, facilitando a navegação e a organização do glossário.

Além disso, o repositório *Canal Ciência* foi utilizado para armazenar e organizar esses dados, permitindo a atualização e a integração contínua das informações. Na Figura 3, é possível observar a estrutura da entrada de cada verbete no repositório.

Figura 2 – Exemplo da entrada do verbete no repositório

Fonte: extraída do repositório *Canal Ciência*.

O gerenciamento da base de dados terminológicos marca a conclusão do processo de construção do glossário. Além disso, o repositório do *Canal Ciência* desempenha um papel crucial ao propiciar a separação entre estrutura de conteúdos e sua apresentação, além de

fornecer a plataforma para o armazenamento e a atualização contínua desses dados, garantindo a sua manutenção e utilização ao longo do tempo.

4 Resultados

Esta seção apresenta os principais resultados obtidos durante o desenvolvimento do glossário, com foco na caracterização do *corpus* utilizado, no processo de validação dos termos técnicos extraídos e na organização conceitual e categorial dos dados terminológicos. Esses resultados refletem a aplicação sistemática da metodologia descrita anteriormente, integrando tecnologias linguísticas, análise semântica e validação colaborativa por especialistas surdos e ouvintes.

4.1 Caracterização do corpus

O *corpus* utilizado neste estudo foi composto por 583 textos científicos coletados automaticamente do portal Canal *Ciência*, serviço de divulgação vinculado ao *Ibict*. Trata-se de uma base textual composta por 323.707 *tokens*, organizados em 15.462 sentenças, provenientes de textos de divulgação científica que abordam temas relacionados à ciência, tecnologia, saúde, meio ambiente, geografia e diversidade populacional.

A Tabela 1 apresenta uma síntese das métricas quantitativas do *corpus*:

Tabela 1 – Caracterização do *corpus*.

Métricas	Quantidade
Total de <i>tokens</i>	323.707
Total de sentenças	15.462
Total de textos	583

Fonte: elaborada pelos autores.

É importante salientar que a composição desse *corpus* foi essencial para garantir um levantamento terminológico consistente, uma vez que os textos apresentam vocabulário técnico acessível ao público geral, porém alinhado aos discursos científicos. Essa característica

o torna particularmente adequado ao propósito de elaboração de um glossário bilíngue voltado à acessibilidade e à inclusão comunicacional da comunidade surda.

4.2 Validação dos termos

A etapa de validação dos termos extraídos do *corpus* constituiu um momento fundamental no processo de construção do glossário, pois buscou assegurar a relevância conceitual, a adequação linguística e a acessibilidade comunicacional dos itens selecionados. Para isso, foram conduzidas duas rodadas de validação, envolvendo especialistas de diferentes áreas e representantes da comunidade surda, respeitando o caráter bilíngue e intermodal do produto final.

4.2.1 Primeira etapa de validação

Na primeira rodada, participaram cinco especialistas ouvintes, todos com experiência e atuação em pesquisa científica. Os especialistas receberam uma lista contendo 1.284 candidatos a termos, previamente extraídos do *corpus* por meio de ferramentas de PLN. A cada especialista foi solicitado avaliar a pertinência dos termos para o domínio da divulgação científica, considerando critérios como clareza, relevância e frequência de uso.

Dessa análise inicial, 98 termos foram validados com índice de concordância mínimo de 80% entre os avaliadores, formando a primeira versão da lista de termos a serem incluídos no glossário. Essa etapa permitiu um primeiro filtro qualitativo, garantindo que apenas termos significativos e representativos das áreas temáticas fossem considerados para o próximo estágio do projeto.

4.2.2 Segunda etapa de validação

Na segunda rodada de validação, o grupo foi ampliado para oito especialistas — seis ouvintes e dois surdos, buscando incorporar a perspectiva da comunidade surda, essencial para garantir que os sinais em Libras propostos fossem cultural e linguisticamente apropriados. Para assegurar essa representatividade, atribuiu-se peso 2 às respostas dos participantes surdos no cálculo de concordância. Mantendo os mesmos 1.284 termos

analisados anteriormente, essa etapa resultou na validação de 130 termos com, no mínimo, 80% de concordância.

Esse processo colaborativo, que integrou vozes surdas e ouvintes, assegurou não apenas o rigor conceitual e linguístico na seleção dos termos, mas também o alinhamento com princípios de acessibilidade e inclusão. Os termos validados foram considerados relevantes, adequados à Libras e formaram a base do glossário bilíngue.

4.3 Resultado da categorização e mapeamento dos termos

Após a validação das 130 unidades lexicais com alto índice de concordância entre os especialistas, os termos foram organizados e classificados segundo critérios semânticos, temáticos e conceituais. Essa etapa teve como objetivo estruturar os dados de maneira sistemática, favorecendo sua usabilidade em um glossário bilíngue voltado à divulgação científica e à acessibilidade linguística em Libras.

A categorização dos termos buscou refletir os principais domínios temáticos abordados nos textos do portal do *Canal Ciência*. A partir da análise dos campos conceituais presentes, foi possível identificar seis *clusters* temáticos centrais, os quais passaram a orientar a organização da base terminológica. São eles:

- **Pesquisa e Divulgação** (ex.: publicação científica, produção acadêmica)
- **Ambiente e Sustentabilidade** (ex.: biodiversidade, desenvolvimento sustentável)
- **Ciência e Tecnologia** (ex.: nanotecnologia, inteligência artificial)
- **Saúde e Genética** (ex.: célula-tronco, biologia molecular)
- **Diversidade e Populações** (ex.: diversidade biológica, inclusão social)
- **Geografia e Urbanização** (ex.: poder público, crescimento urbano)

Cada *cluster* representa um campo semântico que agrupa termos relacionados por afinidade temática, o que facilita tanto a navegação do glossário quanto o entendimento dos conceitos por parte do público-alvo. Após a definição desses *clusters*, os termos foram organizados internamente em subcategorias específicas, conforme nuances de significado ou áreas mais delimitadas dentro do tema central.

Além da categorização, os termos validados foram relacionados entre si por meio de triplas semânticas (sujeito–predicado–objeto), que permitiram a construção de uma rede de

relações conceituais. Por exemplo, triplas como “*pesquisa – é divulgada por – divulgação científica*” ou “*meio ambiente – combate – aquecimento global*” evidenciam conexões semânticas que ajudam a compreender o papel dos termos em um ecossistema de significados. Esse tipo de estrutura contribui para representar de forma lógica e intuitiva as interdependências entre os conceitos, como pode ser observado na Figura 3.

Figura 3 – Exemplo da entrada do verbete no repositório

Pesquisa e Divulgação		
atividade científica	é um	conjunto de práticas
divulgação científica	é um	
avaliação de impacto	é um	processo
método científico	é um	
pesquisa	é um	
pesquisa científica	é um	
popularização	é um	
Ambiente e Sustentabilidade		
aquecimento global	é um	fenômeno
biocombustível	é um tipo de	combustível
biodiesel	é um tipo de	
bioetanol	é um tipo de	
biodegradável	é uma	qualidade

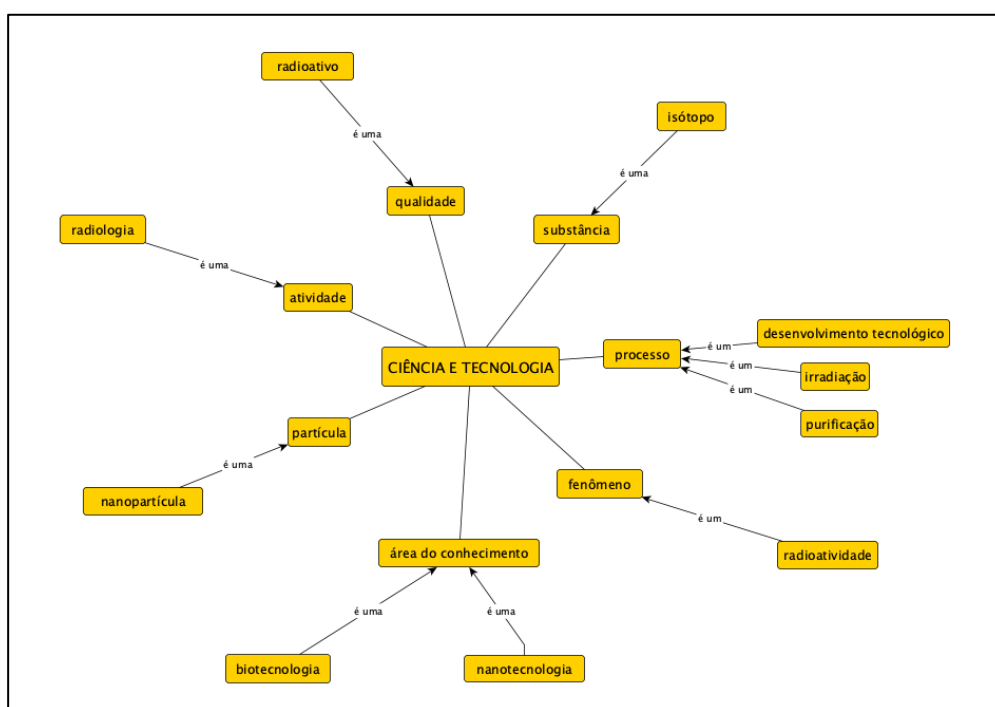
Fonte: elaborada pelos autores.

Assim, a partir dessas triplas, foi possível elaborar uma rede semântica mais ampla, na qual os termos se interligam em múltiplos níveis, formando um mapa conceitual dinâmico e interconectado. Por exemplo, termos como “*base de dados*” e “*armazena*” foram relacionados à “*informação científica*”; já “*projeto de pesquisa*” foi vinculado à “*produção acadêmica*”. Essa organização permite que o glossário atue não apenas como uma lista de termos e definições, mas como uma ferramenta de apoio à compreensão contextual e interdisciplinar.

A criação e a edição do mapa conceitual foram realizadas utilizando a ferramenta *yED Graph Editor*, conhecido por sua interface intuitiva e capacidade de visualização de redes

complexas. A ferramenta possibilitou o agrupamento dos termos validados, a construção gráfica das conexões semânticas e a exportação das representações para integração futura em ambientes digitais, como pode ser observado na Figura 4.

Figura 4 – Exemplo das relações semânticas.



Fonte: elaborada pelos autores.

Portanto, essa etapa de categorização e mapeamento semântico dos termos (Cabré, 1999; Faber, 2009) não apenas estruturou os dados de forma inteligível e navegável, como também favoreceu a posterior etapa de representação bilíngue dos conceitos, assegurando coerência terminológica, consistência temática e potencial pedagógico para usuários ouvintes e surdos.

4.4 Desenvolvimento e gestão da base de dados terminológica

Com a validação final dos 130 termos, foi possível iniciar a construção da base de dados terminológica vinculada ao repositório *Canal Ciência*. Cada termo foi cadastrado com um conjunto de informações essenciais que garantem sua identificação, compreensão e

reutilização. Os campos preenchidos incluíram: forma escrita em português, definição clara e contextualizada, categoria gramatical, contexto de uso, exemplo em português, fonte de validação e domínio ou área temática. A Figura 5 exemplifica como essas informações foram organizadas em uma ficha terminológica no repositório *Canal Ciência*.

Figura 5 – Exemplo da ficha terminológica preenchida no repositório *Canal Ciência*.

Fonte: elaborada pelos autores.

A estrutura proposta para o glossário visa garantir precisão conceitual e acessibilidade, organizando os termos em campos semânticos específicos para facilitar sua compreensão por diferentes públicos. A próxima etapa do projeto envolve a integração da Libras aos termos validados, com a participação conjunta de especialistas surdos e ouvintes para assegurar a adequação cultural e linguística dos sinais. Cada termo será registrado em vídeo e contextualizado nas fichas terminológicas, enquanto a base definicional será finalizada com revisão de definições, exemplos de uso e categorização temática, promovendo a inclusão comunicacional e o acesso democrático ao conhecimento científico.

5 Considerações finais

A metodologia utilizada neste projeto, baseada em técnicas de PLN e embasada pela linguística aplicada e pela educação inclusiva, mostrou-se eficaz na extração, organização e validação de termos científicos relevantes para a construção de um glossário bilíngue Português–Libras. A partir de um *corpus* robusto, composto por textos do portal do *Canal Ciência*, foi possível construir uma base terminológica acessível, alinhada tanto às demandas da comunidade científica quanto às necessidades comunicacionais da população surda.

Entre os principais resultados, destaca-se a validação de 130 termos, obtida por meio de um processo colaborativo que envolveu especialistas falantes e surdos, assegurando rigor técnico e sensibilidade cultural. A categorização desses termos em seis clusters temáticos, combinada à construção de uma rede semântica baseada em triplas (sujeito–predicado–objeto), resultou em uma estrutura conceitual organizada e interconectada, com potencial para facilitar a compreensão dos conceitos e sua aplicação em diferentes contextos de divulgação científica.

Gerir a base terminológica no *Canal Ciência* representa um avanço relevante para a inclusão comunicacional. Integrar vídeos em Libras às fichas enriquecidas amplia o alcance do glossário e reforça o compromisso com uma ciência mais acessível, democrática e sensível às diferenças culturais. Com a inserção contínua de novos termos e o aprimoramento das definições, o glossário torna-se uma ferramenta dinâmica, capaz de acompanhar as transformações do conhecimento científico.

Os resultados aqui apresentados confirmam o potencial das tecnologias linguísticas e do PLN como instrumentos para promover inclusão e reduzir desigualdades no acesso à informação. O trabalho evidencia, ainda, a importância de políticas e práticas que favoreçam a participação ativa da comunidade surda em todas as etapas do processo, desde a concepção até a validação e disseminação do conteúdo.

Como perspectivas futuras, recomenda-se a ampliação do glossário para contemplar outras áreas do conhecimento científico, bem como o fortalecimento de ações colaborativas com instituições de ensino, pesquisa e associações de surdos. A consolidação dessa iniciativa

poderá inspirar novos projetos voltados à acessibilidade linguística e à equidade na produção e circulação do saber.

5.1 Aspectos Éticos

Os autores declaram que a presente pesquisa não foi submetida ao Comitê de Ética em Pesquisa (CEP) por não envolver riscos aos participantes nem coleta de dados sensíveis. O estudo garantiu anonimato, confidencialidade e não incluiu identificação ou exposição dos informantes. Não houve intervenções diretas junto aos participantes. Ainda assim, foram respeitados os princípios éticos que orientam a pesquisa científica.

Referências

ANTHONY, L. AntConc (Versão 4.2.2) [Software]. Tóquio, Japão: Waseda University, 2023. Disponível em: <https://www.laurenceanthony.net/software>. Acesso em: 12 jun. 2025.

BJÖRK, Bo-Christer. Open access to scientific publications – an analysis of the barriers to change? **Information Research**, v. 10, n. 2, 2005. Disponível em: <https://informationr.net/ir/10-2/paper170.html>. Acesso em: 10 jun. 2024.

BRASIL. Decreto nº 5.296, de 2 de dezembro de 2004. Regulamenta as Leis nº 10.048/2000 e nº 10.098/2000, que estabelecem normas gerais e critérios básicos para a promoção da acessibilidade. **Diário Oficial [da] República Federativa do Brasil**, Brasília, DF, 3 dez. 2004.

BUENO, W. S. **Jornalismo científico no Brasil: os compromissos de uma prática dependente**. São Paulo: Unesp, 2010.

CABRÉ, M. T. **Terminology: Theory, Methods, and Applications**. Edited by Juan C. Sager. Illustrated edition. Amsterdam: John Benjamins Publishing, 1999. 247 p. (Terminology and Lexicography Research and Practice, v. 1).

DIAS-DA-SILVA, B. C. O estudo linguístico-computacional da linguagem. **Letras de Hoje**, v. 41, n. 2, p. 103-138, 2006. Disponível em: <https://revistaseletronicas.pucrs.br/ojs/index.php/fale/article/view/597>. Acesso em: 10 jun. 2025.

DI FELIPPO, A.; ALMEIDA, G. M. Uma metodologia para o desenvolvimento de Wordnets terminológicas em português do Brasil. In: MOITA LOPES, L. P.; ALMEIDA FILHO, F. M. (Org.). **Estudos em linguística computacional: teorias e aplicações**. São Paulo: Mercado de Letras, 2010. p. 365-395. DOI: <https://doi.org/10.11606/issn.2317-9511.tradterm.2010.46325>.

FABER, Pamela. The cognitive shift in terminology and specialized translation. **Monografías de Traducción e Interpretación – MonTI**, n. 1, p. 107–134, 2009.

LOPES, L.; VIEIRA, R.; FINATTO, M. J. B.; ZANETTE, A.; MARTINS, D.; RIBEIRO JUNIOR, L. C. Extração automática de termos compostos para construção de ontologias: um experimento na área da saúde. **RECIIS: revista eletrônica de comunicação, informação & inovação em saúde**, Rio de Janeiro, vol. 3, n. 1, p. 79-88, 2009.

MASSARANI, L.; MOREIRA, I. C.; BRITO, F. **Ciência e público**: caminhos da divulgação científica no Brasil. Rio de Janeiro: Casa da Ciência/UFRJ, 2002. p. 185-202.

OLIVEIRA, R. L. **A educação dos surdos**: avanços e desafios. 2015. Dissertação (Mestrado em Educação) – Universidade Federal de Ouro Preto, Ouro Preto, 2015.

PROMETI, D.; COSTA, M. R.; TUXI, P. Sinal-termo, língua de sinais e glossário bilíngue: atuação da universidade de Brasília nas pesquisas terminológicas. *In*: CONGRESSO NACIONAL DE LIBRAS DA UNIVERSIDADE FEDERAL DE UBERLÂNDIA, 1., 2015, Uberlândia. **ANAIS I CONALIBRAS-UFU**. Uberlândia: s.e., 2015.

SANTOS, H. R. O estudo da Terminologia em Libras com o apoio de Corpora. **Revista Diálogos**, v. 6, n. 1, p. 61-70, 2018.

STROBEL, K. **As imagens do outro sobre a cultura surda**. Florianópolis: Editora da UFSC, 2008.

TUXI, P. Proposta de organização de verbete em Glossários terminológicos bilíngues – língua Brasileira de sinais e língua portuguesa. **Cad. Trad.**, Florianópolis, v. 35, nº especial 2, p. 557-588, jul.-dez. 2015.

Recebido em: 15.07.2025

Aprovado em: 01.06.2026