GLOSSÁRIOS MÉDICOS MULTILÍNGUES PARA *CAT TOOLS*: DESENVOLVIMENTO BASEADO NA CID-11 DA OMS

MULTILINGUAL MEDICAL GLOSSARIES FOR CAT TOOLS: DEVELOPMENT BASED ON WHO ICD-11

DOI: 10.14393/LL63-v41S-2026-02

Jean-Claude Miroir*

RESUMO: A tradução médica enfrenta desafios relacionados à consistência terminológica e à ausência de recursos multilíngues padronizados. Estudos em Terminologia e Tradução Especializada demonstram que glossários estruturados aumentam a precisão em traduções técnicas quando integrados a *CAT tools*. Apesar da disponibilidade da Classificação Internacional de Doenças (CID-11) da Organização Mundial da Saúde, seus dados não estão estruturados em formatos compatíveis. Este artigo apresenta uma metodologia para transformar dados oficiais da CID-11 em glossários multilíngues funcionais. Foram compilados 36.049 pares terminológicos em quatro idiomas (inglês, português, francês e espanhol), estruturados em formatos TBX e XLSX, assegurando controle de qualidade no processo tradutório. Os dados possibilitam exploração linguística por meio de concordanciadores, cujo uso requer enriquecimento mediante processamento de linguagem natural. Etiquetadores morfossintáticos (spaCy e *Stanza*) revelaram limitações dos modelos generalistas na classificação de nomes próprios, com taxa de erro superior a 95%. Os recursos foram disponibilizados no Figshare.

PALAVRAS-CHAVE: Tradução médica. Terminologia multilíngue. Processamento de linguagem natural. Classificação Internacional de Doenças. *CAT Tools*.

ABSTRACT: Medical translation faces challenges related to terminological consistency and the absence of standardized multilingual resources. Studies in Terminology and Specialized Translation demonstrate that structured glossaries increase accuracy in technical translations when integrated into CAT tools. Despite the availability of the World Health Organization's International Classification of Diseases (ICD-11), its data are not structured in compatible formats. This article presents a methodology to transform official ICD-11 data into functional multilingual glossaries. A total of 36,049 terminological pairs were compiled in four languages (English, Portuguese, French, and Spanish), structured in TBX and XLSX formats, ensuring quality control in the translation process. The data enable linguistic exploration through concordancers, whose use requires enrichment through natural language processing. Part-of-speech taggers (spaCy and *Stanza*) revealed limitations of generalist models in classifying proper nouns, with an error rate exceeding 95%. The resources were made available on Figshare.

KEYWORDS: Medical translation. Multilingual terminology. Natural language processing. International Classification of Diseases. *CAT Tools*.

* Doutor em Literatura. Universidade de Brasília. ORCID: Graduanda em Tradução – Inglês pela Universidade de Brasília. ORCID: 0000-0002-5875-9074. E-mail: jc.unb@.com.

1 Introdução

A tradução médica constitui um domínio de alta especialização em que a precisão terminológica pode determinar a segurança dos pacientes e a eficácia da comunicação científica internacional. Stefaniak (2017) enfatiza que a gestão terminológica é parte integral de todo processo tradutório e necessária para alcançar traduções de alta qualidade, destacando que a integração de terminologia no processo tradutório garante consistência, precisão e clareza.

Em um cenário caracterizado pela complexidade de nomenclaturas como a CID (Classificação Internacional de Doenças), SNOMED CT (*Systematized Nomenclature of Medicine – Clinical Terms*) (SNOMED, 2025) e outros sistemas classificatórios, tradutores da área da saúde enfrentam desafios significativos para manter a consistência terminológica e evitar ambiguidades que possam comprometer a qualidade. Fung *et al*. (2024) exploraram o mapeamento entre SNOMED CT e a CID-11 para melhorar a interoperabilidade semântica entre dois importantes sistemas de terminologia em saúde.

A Organização Mundial da Saúde (OMS) desempenha papel fundamental na padronização terminológica global por meio de suas classificações internacionais. Prieto Ramos (2020) enfatiza o papel crucial dos recursos terminológicos em organizações internacionais para assegurar consistência e precisão na tradução de termos especializados, destacando que terminologias padronizadas reduzem ambiguidades e aumentam a qualidade das traduções institucionais.

A transição da CID-10 para a CID-11, implementada globalmente a partir de 2022, representa um marco na modernização dos sistemas de informação em saúde, baseado em ontologia formal (Harrison, J. E.; Weber, S.; Jakob, R.; Chute, C. G., 2021). Enquanto a CID-10 foi desenvolvida originalmente para o formato impresso, limitando sua interoperabilidade digital, a CID-11 foi concebida desde o início como ferramenta eletrônica, otimizada para integração com sistemas informatizados como Prontuários Eletrônicos de Saúde (PES) e *CAT Tools*. Esta evolução tecnológica cria oportunidades sem precedentes para o desenvolvimento de recursos terminológicos avançados. O processamento de linguagem natural (PLN) tornou-se uma tecnologia-chave para automatizar a extração, o alinhamento e a estruturação de dados

terminológicos multilíngues, possibilitando a criação de glossários especializados com precisão e escalabilidade.

Essa aplicação do projeto Tradução Médica em Contexto (TraMeC)[1] propõe uma metodologia inovadora que combina dados terminológicos oficiais da CID-11 com tecnologias de PLN (*spaCy* e *Stanza*) para desenvolver ferramentas terminológicas de alta precisão. A inovação não se encontra nas ferramentas, mas na ponte metodológica entre recursos terminológicos massivos, raramente explorados por tradutores, e aplicação prática em contextos profissionais. Assim, o objetivo central deste artigo consiste em transformar os recursos multilíngues disponibilizados pela OMS em glossários funcionais no formato TBX (*TermBase eXchange*), padrão internacional ISO 30042 (International Organization for Standardization, 2019a) compatível com as principais *CAT Tools*. A metodologia desenvolvida, fundamentada em Linguística de Corpus (LC) e PLN, processa mais de 36.000 descrições de doenças em inglês, estabelecendo alinhamentos precisos com suas traduções oficiais para português, francês e espanhol. Este corpus terminológico paralelo, validado pela OMS, permite a criação de glossários nos formatos TBX e XLSX, compatíveis com ferramentas como *Wordfast Pro* (Champollion, 2025) ou *Matecat* (Fondazione Bruno Kessler, 2025). A disponibilização destes recursos no *Figshare* (Miroir, 2025) democratiza o acesso à terminologia médica oficial, devidamente organizada em português, e oferece à comunidade tradutória ferramentas padronizadas e adaptáveis.

Este trabalho apresenta dupla contribuição: metodológica, por meio do desenvolvimento de um *pipeline* replicável para criação de recursos terminológicos multilíngues; e prática, mediante a disponibilização de glossários funcionais que podem elevar a qualidade e consistência das traduções médicas. Assim, esta pesquisa investiga as potencialidades e limitações das tecnologias de PLN atuais quando aplicadas à terminologia médica altamente especializada, contribuindo para o avanço do conhecimento na interface entre linguística computacional e tradução especializada. Esta iniciativa visa incentivar a adoção

[1] O projeto Tradução Médica em Contexto (TraMeC) é um projeto de pesquisa interdisciplinar dedicado ao estudo aprofundado da tradução e da comunicação especializada no domínio da saúde. Seu objetivo principal é analisar e otimizar os processos de transposição linguística e cultural de conteúdos médicos, garantindo precisão, acessibilidade e eficácia na disseminação de informação clínica e científica entre profissionais, pacientes e diferentes contextos socioculturais, como o desenvolvimento de glossários e recursos terminológicos para apoiar tradutores e profissionais da saúde.

de práticas padronizadas na tradução médica, promovendo maior confiabilidade nas traduções e comunicação mais precisa entre diferentes sistemas de saúde globais. O projeto TraMeC busca estabelecer um modelo escalável e sustentável para o desenvolvimento de recursos terminológicos especializados, aplicável a outras áreas que demandem precisão terminológica similar, como Prontuários Eletrônicos de Saúde (PES ou *Electronic Health Records* - EHR).

Na era da IA, essa convergência entre recursos oficiais massivos, tecnologias de PLN e necessidades práticas de tradutores representa uma abordagem inovadora, estabelecendo procedimentos replicáveis que democratizam o acesso à terminologia médica padronizada e possibilitam análises linguísticas avançadas. A contribuição original é tornar dados oficiais da OMS amplamente funcionais para tradutores profissionais.

2 Apresentação dos dados

A OMS oferece aos usuários, em seu site, uma página inicial sobre a Classificação Internacional de Doenças, 11ª Revisão (doravante CID-11), que apresenta *"*O padrão global para informações de diagnóstico de saúde*"* (OMS, 2025a). A seção *"*Use a CID-11*"* disponibiliza ao usuário o Navegador da *"*CID-11 para Estatísticas de Mortalidade e de Morbidade*"* (doravante CID-11 MMS[2]) (OMS, 2025b), uma ferramenta desenvolvida pela OMS para padronizar a codificação de causas de morte e condições de morbidade, seguindo os padrões da CID-11. Essa ferramenta oferece uma interface para busca de códigos, facilitando a localização de categorias específicas de saúde.

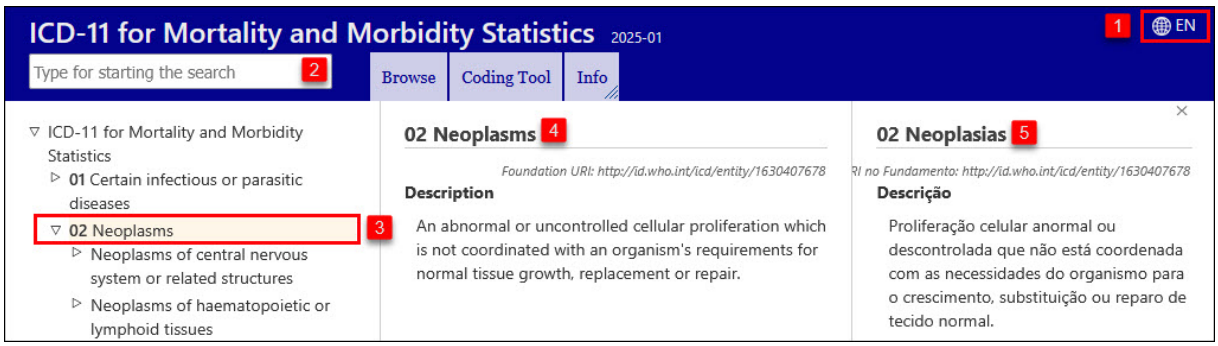
Com funcionalidades para integração a sistemas de informação de saúde e materiais de capacitação, o navegador CID-11 MMS é amplamente utilizado por uma variedade de profissionais e organizações no campo da saúde. Isso inclui médicos e enfermeiros que codificam diagnósticos, epidemiologistas que analisam padrões de doenças, pesquisadores que conduzem estudos científicos, gestores de saúde pública que planejam e avaliam programas de saúde e tradutores profissionais da área da saúde. Além disso, instituições de saúde, como hospitais e clínicas, bem como organizações internacionais e governamentais, incluindo a OMS

[2] A sigla "MMS" refere-se, em inglês, a "*Mortality and Morbidity Statistics.*"

e ministérios da saúde, utilizam essa ferramenta para garantir a coleta, codificação e análise consistente de dados de saúde em diferentes contextos e regiões do mundo.

O navegador CID-11 MMS[3] está disponível, no momento, em 14 idiomas: árabe, cazaque, chinês, eslovaco, espanhol, francês, inglês, latim, português, russo, sueco, tcheco, turco e uzbeque. Além disso, o mesmo navegador permite abrir uma coluna adicional para visualizar as informações em duas línguas em paralelo, conforme apresentado a seguir:

Figura 1 – Aspectos do navegador CID-11 MMS em configuração bilíngue: inglês - português



Fonte: extraída da página do navegador CID-11 MMS (OMS, 2025b)

Ao clicar no ícone de idioma [1] (Figura 1), os usuários podem escolher um par de idiomas: um principal e um secundário, para fins de comparação. Após selecionar os idiomas, ao pesquisar um termo específico [2] e selecionar um item do menu principal [3], o navegador exibirá as informações no idioma principal [4], no exemplo, em inglês. Simultaneamente, ele mostrará a descrição correspondente no idioma secundário [5], neste caso, o português.

O navegador CID-11 MMS oferece aos usuários da área da saúde acesso simultâneo a informações detalhadas em duas línguas. Para os profissionais, tradutores da área da saúde, essa apresentação em paralelo oferece inúmeras vantagens, como o esclarecimento sobre o sentido correto dos termos em contexto. No entanto, a consulta repetitiva ao navegador pode se tornar demorada e incluir níveis de detalhe que nem sempre são úteis para fazer escolhas terminológicas ou conceituais precisas.

[3] ICD-11 for Mortality and Morbidity Statistics – Browser. Disponível em: https://icd.who.int/browse/2025-01/mms/en . Acesso em: 20 abr. 2025.

Assim, esse estudo tem como objetivo principal realizar o alinhamento terminológico preciso entre termos médicos e suas traduções oficiais para compilar um glossário especializado. O método previa a criação de um script em Python para automatizar a extração desses pares, uma estratégia necessária para a construção desse tipo de base terminológico como no caso da CID-10.

Todavia, para a nova CID-11, a OMS disponibilizou esse alinhamento de forma estruturada. Esse avanço viabiliza o uso dos dados em ferramentas computacionais, como Prontuários Eletrônicos do Paciente (PEP) ou *CAT Tools* (*Wordfast Pro* ou *Matecat*), facilitando a implementação desses recursos de padronização terminológica.

Fernández-Parra (2025) aborda questões fundamentais sobre gestão terminológica para tradutores, incluindo o nível de conhecimento necessário sobre terminologia e seu gerenciamento, além das ferramentas e tecnologias que podem auxiliar tradutores na gestão de terminologia especializada e apresenta uma visão atualizada sobre práticas de gestão terminológica baseada em dados empíricos. Os usuários interessados podem encontrar os dados, no Menu da aba Info (Figura 1, acima) do navegador CID-11 MMS, um *link* chamado "Arquivo em Planilha" (na versão em português). Esse *link* disponibiliza para download o arquivo compactado *SimpleTabulation-ICD-11-MMS-pt.zip*, que contém três arquivos:

1) *readme.txt*: Esse arquivo é o *"*dicionário de dados*"* que descreve os campos (estrutura, tipos de dados, relações e outras informações) de um banco de dados.
2) *SimpleTabulation-ICD-11-MMS-pt.txt*: O arquivo utiliza o formato TSV (*Tab-Separated Values*)[4], onde os dados são organizados em linhas (registros) e colunas (campos das características dos dados) separadas por tabulações.
3) *SimpleTabulation-ICD-11-MMS-pt.xlsx*: Assim como o TSV, o formato XLSX, de planilhas *Excel*, mantém a estrutura básica de linhas (registros) e colunas (campos), mas acrescenta recursos avançados como formatação visual, filtros e fórmulas, tornando a análise interna mais acessível para os usuários.

[4] Essa estrutura é ideal para bases tabulares, pois evita conflitos com vírgulas, comum em arquivos CSV (*Comma - Separated Values*), mantendo a simplicidade de um arquivo de texto, porém com separação mais robusta para dados que contenham pontuação.

A análise criteriosa dos dados permite destacar duas colunas com os nomes dos campos *TitleEN* (título em inglês) e *Title* (título na língua-alvo). A inclusão desses dois campos nesse banco de dados (planilha) é estratégica para o desenvolvimento de ferramentas de apoio à tradução, com a terminologia oficial da CID-11, cumprindo funções complementares essenciais.

Além dos benefícios para tradutores, essa estrutura em paralelo auxilia os profissionais de saúde a navegarem entre as versões linguísticas, promovendo melhor entendimento das variações diagnósticas em diferentes idiomas, o que é fundamental para a interoperabilidade internacional de sistemas de saúde.

A análise do banco de dados disponível no arquivo *SimpleTabulation-ICD-11-MMS-pt.xlsx* revela que ele contém 36.782 entradas terminológicas bilíngues organizadas, conforme o trecho seguinte do Quadro 1:

Quadro 1 – Trecho com entradas terminológicas bilíngues[5]

	E	F
N°	TitleEN	Title
1	*Certain infectious or parasitic diseases*	Algumas doenças infecciosas ou parasitárias
2	- *Gastroenteritis or colitis of infectious origin*	- Gastroenterites ou colites de origem infecciosa
3	- - *Bacterial intestinal infections*	- - Infecções intestinais bacterianas
4	- - - *Cholera*	- - - Cólera
5	- - - *Intestinal infection due to other Vibrio*	- - - Infecção intestinal por outras bactérias do gênero Vibrio

Fonte: elaborado pelo autor.

Observe-se no Quadro 1 que os dados terminológicos bilíngues são organizados em duas colunas justapostas, intituladas *"TitleEN"* (coluna E), para os termos em inglês, e apenas *"Title"* (coluna F), para os termos em português. Essa disposição facilita significativamente a extração e a exploração desses dados em glossários, conforme será descrito no próximo item de Metodologia. A mesma estrutura é observada nos arquivos coletados: (1) traduções em francês, *SimpleTabulation-ICD-11-MMS-fr.xlsx*, com as colunas *"TitleEN"* e *"Title"* (termos em francês), e (2) traduções em espanhol, *SimpleTabulation-ICD-11-MMS-es.xlsx*, com as colunas *"TitleEN"* e *"Title"* (termos em espanhol). A coluna *"TitleEN"* desempenha um papel fundamental como âncora para o alinhamento preciso das demais traduções.

[5] As colunas E e F do trecho da planilha correspondem aos cabeçalhos TitleEN e Title, exibindo os títulos das categorias diagnósticas em inglês e português, respectivamente.

No entanto, na prática, traduções médicas são frequentemente realizadas em pares de línguas que não envolvem o inglês, como francês-português ou espanhol-português. Nesse contexto, é possível organizar o alinhamento terminológico com precisão, utilizando o inglês como âncora. Dessa forma, pode-se extrair posteriormente os pares de línguas desejados, sem a necessidade de incluir o inglês, conforme será descrito no próximo item.

3 Abordagem simplificada da extração bilíngue com inglês como âncora

A abordagem simplificada[6] oferece um método intuitivo que não requer domínio avançado de Excel ou Python para manipulação de dados linguísticos, permitindo a criação e gestão eficiente de recursos terminológicos bilíngues. A versão original adotou abordagem *hands-on*, apresentando etapas metodológicas de forma descritiva para facilitar a replicação. Nesse formato, as justificativas emergiam implicitamente durante a execução dos procedimentos.

3.1 Preparação do material

O primeiro passo da preparação do material é criar um novo arquivo Excel (.xlsx) bilíngue nomeado *ICD-11-MMS-en-pt.xlsx*, que armazenará exclusivamente os termos em inglês e suas traduções correspondentes em português, sem incluir outros metadados. Em seguida, o arquivo original da OMS, intitulado *SimpleTabulation-ICD-11-MMS-pt.xlsx*, deve ser acessado para copiar as colunas E (*TitleEN* - contendo os termos em inglês) e F (*Title* - com as respectivas traduções em português). Na primeira planilha do novo arquivo *ICD-11-MMS-en-pt.xlsx*, os dados copiados[7] devem ser inseridos (“colados”) nas colunas A (termos em inglês) e B (traduções para português).

Observa-se que o *Módulo II* do Item *“26 - Capítulo Suplementar: Condições da Medicina Tradicional”* ainda não foi traduzido pela OMS para o português, nem para as demais línguas disponíveis. Do total de 36.782 termos em inglês contidos no glossário, identificaram-se 733

[6] "Simplificada" em comparação com a CID-10, que exige *scripts* Python para extração e alinhamento multilíngue. Na CID-11, a OMS disponibilizou planilha Excel estruturada (*SimpleTabulation-ICD-11-MMS*), reduzindo a complexidade técnica e democratizando o acesso aos dados oficiais.

[7] Recomenda-se a remoção dos marcadores hierárquicos de tipo [- - - -] (sem os colchetes) que precedem cada termo (entrada), por se mostrarem doravante desnecessários.

itens sem tradução correspondente, resultando em 36.049 pares de termos alinhados (inglês-português). Para manter a integridade do glossário, os 733 termos não traduzidos foram sistematicamente extraídos do arquivo *ICD-11-MMS-en-pt.xlsx*, eliminando, assim, a presença de células vazias que poderiam comprometer a eficiência dos recursos terminológicos criados. Esses termos são armazenados em um novo arquivo *ICD-11-MMS-TM2-en-pt.xlsx* e serão traduzidos, nas próximas etapas, com o auxílio dos glossários TBX e XLSX que foram criados. Todos os termos não traduzidos são associados à sigla *"*(TM2)*"* (*Traditional Medicine* do *Module II*). Para ser traduzida, a coluna A foi extraída e copiada em um arquivo .txt simples e renomeado *ICD-11-MMS-TM2-en.txt*

3.2 Tipos de conversões operacionais em CAT Tools

As ferramentas de gestão de terminologia ajudam a padronizar a tradução de termos especializados, em áreas de conhecimento que exigem exatidão, como a médica, jurídica e tecnológica, evitando ambiguidades causadas pelo uso inconsistente das unidades terminológicas. Schoening (2024) destaca que as bases terminológicas (*term bases*) desempenham um papel fundamental na manutenção da precisão tradutória em múltiplos projetos, possibilitando o uso consistente de terminologia especializada relevante para a organização. Entre os modelos disponíveis, destacam-se duas abordagens complementares: (1) o padrão TBX (*TermBase eXchange*) para garantir a interoperabilidade entre sistemas de tradução; e (2) modelos de planilhas eletrônicas, XLSX ou ODS[8], adaptáveis às exigências específicas de gestão terminológica das *CAT Tools*, que podem não usar o formato TBX.

3.2.1 Padrão TBX

A primeira modalidade de conversão terminológica implementada consiste na transformação do glossário da OMS para o padrão TBX (*TermBase eXchange*). Esse padrão é baseado em XML[9] e foi desenvolvido para a troca de dados terminológicos entre sistemas, garantindo a interoperabilidade entre *CAT Tools*, como *SDL Trados* (RWS Group, 2025), *MemoQ*

[8] Open Document Spreadsheet, formato aberto do LibreOffice/OpenOffice.
[9] XML (eXtensible Markup Language) é uma linguagem de marcação padronizada (W3C) que permite estruturar dados de forma hierárquica e legível por máquina.

(Kilgray Translation Technologies, 2025), entre outras. Esse formato, regulado pela norma ISO 30042 (International Organization for Standardization, 2019a), permite a estruturação de termos, facilitando a padronização em projetos de tradução e localização, em conformidade com a ISO/TC 37 (International Organization for Standardization, 2019b). O TBX é amplamente utilizado na indústria de tradução para manter a consistência terminológica, especialmente em domínios técnicos, mas também em trabalhos colaborativos, onde múltiplos tradutores atuam simultaneamente em um mesmo projeto de tradução.

Esta primeira modalidade de conversão foi implementada mediante a transformação dos dados extraídos do arquivo *ICD-11-MMS-en-pt.xlsx* para o formato TBX. O processo foi executado com o software *Goldpan* (2021), um editor multifuncional especializado em formatos TBX e TMX com a capacidade de conversão bidirecional entre formatos de glossários e memória de tradução (XLSX, TBX, TMX, XLIFF), além de outras funcionalidades.

3.2.2 Planilha XLSX específica da CAT Tool

A segunda modalidade de conversão terminológica implementada transforma o glossário da OMS em uma planilha Excel, seguindo o modelo padrão (*template*) utilizado pelo *Matecat*, uma vez que esta CAT Tool, escolhida para o estudo, não aceita o formato TBX. Com efeito, o *Matecat* é uma *CAT Tool on-line* gratuita e de código aberto, para tradutores autônomos, estudantes e pesquisadores em tradução ou empresas de tradução. É um projeto acadêmico colaborativo, desenvolvido por instituições de pesquisa e universidades, incluindo o centro internacional FBK (Fondazione Bruno Kessler), a empresa de tradução *Translated srl*, a Université du Maine (França) e a Universidade de Edimburgo (Reino Unido).

O arquivo Excel do modelo padrão do *Matecat* é renomeado para *TB_ICD-11-MMS-en-pt.xlsx*. O prefixo *"*TB*"* identifica o tipo de conteúdo, neste caso, uma *TermBase* (glossário). Essa distinção é necessária, principalmente no *Matecat*, para evitar confusão com a memória de tradução (*Translation Memory*), como no arquivo *TM_ICD-11-MMS-en-pt.tmx* (identificação com o prefixo "TM").

Os dados são extraídos do arquivo *ICD-11-MMS-en-pt.xlsx*, preparado na etapa anterior. Dessa forma, os dados da coluna A *"TitleEN"* são copiados para a coluna E *"*en-US*"*, enquanto

os da coluna B *"*Title*"* são copiados para a coluna H *"*pt-BR*"*. O arquivo está doravante pronto para ser importado no *Matecat*.

Este procedimento, conforme o Quadro 2, é replicado para os pares linguísticos en-es (inglês-espanhol) e en-fr (inglês-francês), modificando-se apenas os códigos de idioma nas colunas E e H.

Quadro 2 – Síntese da organização dos arquivos manipulados e criados neste item[10]

A) Fonte original da OMS *(com metadados)*	
SimpleTabulation-ICD-11-MMS-pt.xlsx	
SimpleTabulation-ICD-11-MMS-fr.xlsx	
SimpleTabulation-ICD-11-MMS-es.xlsx	
▼	
B) Compilação da terminologia bilíngue *(sem metadados, sem níveis, sem vazios)*	
ICD-11-MMS-en-pt.xlsx	
ICD-11-MMS-en-fr.xlsx	
ICD-11-MMS-en-es.xlsx	
▼	
C) Glossários para *CAT Tools*	
C1) TBX gerado com *Goldpan*	**C2) Template XLSX do *Matecat***
TB_ICD-11-MMS-en-pt.tbx	TB_ICD-11-MMS-en-pt.xlsx
TB_ICD-11-MMS-en-fr.tbx	TB_ICD-11-MMS-en-fr.xlsx
TB_ICD-11-MMS-en-es.tbx	TB_ICD-11-MMS-en-es.xlsx

Fonte: elaborado pelo autor.

4 Abordagem simplificada da extração multilíngue sem o inglês

Este método de extração multilíngue sem o inglês mantém a simplicidade da abordagem anterior, oferecendo uma forma intuitiva de criar e gerenciar ferramentas de terminologia multilíngue. Além disso, atende à necessidade de traduzir diretamente entre pares de línguas sem a obrigatoriedade de incluir o inglês como intermediário, por exemplo, os pares es-pt, pt-fr ou fr-es.

[10] Os recursos criados nesta pesquisa encontram-se disponíveis no repositório de acesso aberto *Figshare* (Miroir, 2025)

4.1 Preparação do material

O arquivo original da OMS, explorado anteriormente, nomeado *SimpleTabulation-ICD-11-MMS-pt.xlsx* (arquivo base), deve ser agora duplicado. A cópia pode ser renomeada como *SimpleTabulation-ICD-11-MMS-en-pt-fr-es.xlsx,* para incluir as traduções em francês e espanhol,[11] alinhadas com os termos em inglês e português (já contidos no arquivo base). Para isso, é preciso baixar do site da OMS os arquivos *SimpleTabulation-ICD-11-MMS-fr.xlsx* e *SimpleTabulation-ICD-11-MMS-es.xlsx*. A preparação do arquivo *SimpleTabulation-ICD-11-MMS-en-pt-fr-es.xlsx* consiste em inserir duas colunas ao lado direito da coluna F *"Title"* (renomeada em *"TitlePT"*): a coluna G *"TitleFR"* e a coluna H *"TitleES"*. Esse procedimento é essencial para garantir o alinhamento preciso entre os termos nos diferentes idiomas e as futuras explorações dos metadados para eventual enriquecimento de glossário, conforme o trecho apresentado no Quadro 3, seguinte:

Quadro 4 – Trecho com entradas terminológicas multilíngues

	E	F	G	H
N°	TitleEN	TitlePT	TitleFR	TitleES
1	Certain infectious or parasitic diseases	Algumas doenças infecciosas ou parasitárias	Certaines maladies infectieuses ou parasitaires	Algunas enfermedades infecciosas o parasitarias
2	- Gastroenteritis or colitis of infectious origin	- Gastroenterites ou colites de origem infecciosa	- Gastroentérite ou colite d'origine infectieuse	- Gastroenteritis o colitis de origen infeccioso
3	- - Bacterial intestinal infections	- - Infecções intestinais bacterianas	- - Infections intestinales bactériennes	- - Infecciones intestinales bacterianas
4	- - - Cholera	- - - Cólera	- - - Choléra	- - - Cólera
5	- - - Intestinal infection due to other Vibrio	- - - Infecção intestinal por outras bactérias do gênero Vibrio	- - - Infection intestinale due à d'autres Vibrio	- - - Infección intestinal por otro Vibrio

Fonte: elaborado pelo autor.

Considerando esse corpus terminológico paralelo, em quatro línguas (EN, PT, FR, ES) e excluindo o inglês, é possível estabelecer três pares de tradução indireta (o inglês como língua

[11] Podem ser incluídos os demais idiomas disponíveis em colunas adjacentes: árabe, chinês, tcheco, cazaque, latim, russo, eslovaco, sueco, turco, uzbeque.

intermediária ou pivô): (1) português e francês (PT <> FR), (2) português e espanhol (PT <> ES) e (3) francês e espanhol (FR <> ES).

4.2 Preparação das combinações bilíngues

De forma similar às seções anteriores, o primeiro passo consiste em criar um arquivo Excel (.xlsx) bilíngue contendo apenas o par de idiomas desejado (por exemplo, PT <> FR), que pode ser nomeado como *ICD-11-MMS-pt-fr.xlsx*. Esse arquivo armazenará exclusivamente os termos em português e suas respectivas traduções em francês, excluindo quaisquer outros metadados. Em seguida, o arquivo da OMS, devidamente preparado e intitulado *SimpleTabulation-ICD-11-MMS-en-pt-fr-es.xlsx*, deve ser acessado para extração das colunas E (*TitlePT* - contendo os termos em português) e F (*TitleFR* - com as respectivas traduções em francês). Na primeira planilha do *ICD-11-MMS-pt-fr.xlsx*, os dados extraídos devem ser inseridos nas colunas A (termos em português) e B (termos em francês).

4.3 Tipos de conversões operacionais em CAT Tools

Conforme especificado no item 3.2 , implementam-se também duas soluções complementares: (1) o padrão TBX, com o arquivo nomeado *TB_ ICD-11-MMS-pt-fr.tbx*, e (2) o formato XLSX ou ODS, com o arquivo *TB_ICD-11-MMS-pt-fr.xlsx*, para integração com ferramentas de tradução, como o *Matecat*. Essa plataforma suporta glossários com até 10 idiomas, permitindo combinações linguísticas bidirecionais entre eles: Língua A <> Língua B. A organização sistemática dos arquivos multilíngues criados nesta etapa encontra-se sintetizada no Quadro 4, seguinte:

Quadro 5 - Síntese da organização dos arquivos multilíngues criados neste item[12]

A) Fonte original da OMS *(com metadados)*
SimpleTabulation-ICD-11-MMS-pt.xlsx
SimpleTabulation-ICD-11-MMS-fr.xlsx
SimpleTabulation-ICD-11-MMS-es.xlsx
▼

[12] Os recursos criados nesta pesquisa encontram-se disponíveis no repositório de acesso aberto Figshare (Miroir, 2025).

B) Alinhamento multilíngue da terminologia *(com metadados)*	
SimpleTabulation-ICD-11-MMS-en-pt-fr-es.xlsx	
▼	
C) Compilação da terminologia bilíngue *(sem metadados, sem vazios, sem níveis)*	
ICD-11-MMS-pt-fr.xlsx	
ICD-11-MMS-pt-es.xlsx	
ICD-11-MMS-es-fr.xlsx	
▼	
D) Glossários para *CAT Tools*	
D1) TBX gerado com *Goldpan*	**D2) *Template* XLSX do *Matecat***
TB_ICD-11-MMS-pt-fr.tbx	TB_ICD-11-MMS-pt-fr.xlsx
TB_ICD-11-MMS-pt-es.tbx	TB_ICD-11-MMS-pt-es.xlsx
TB_ICD-11-MMS-es-fr.tbx	TB_ICD-11-MMS-es-fr.xlsx

Fonte: elaborado pelo autor.

5 Implementação em *CAT Tools*

O objetivo desse tópico é apresentar o processo de tradução com o apoio das ferramentas terminológicas criadas em: (1) programas de tradução instalados no computador do usuário (*stand alone*) e (2) em plataforma de tradução *on-line*. Duas implementações foram desenvolvidas, neste artigo, para testar diferentes formatos de glossário. Na primeira, utilizou-se a versão de teste do *Wordfast Pro* que oferece suporte nativo ao formato TBX. Na segunda abordagem, elaborou-se um glossário em formato Excel adaptado às exigências do sistema *on-line Matecat*.

5.1 Wordfast Pro

A versão de teste do *Wordfast Pro* (Champollion, 2025) oferece um período de avaliação com funcionalidades completas, permitindo aos usuários experimentarem a ferramenta em projetos reais de tradução. Durante esse período, o software opera com algumas limitações técnicas: as memórias de tradução ficam restritas a 500 unidades de tradução (UT), há acesso a apenas um glossário, não há integração com tradução automática e não é possível conectar-se a memórias de tradução e glossários remotos. Apesar dessas restrições, a versão de demonstração mantém os recursos essenciais da ferramenta, incluindo filtragem avançada de segmentos, controle de qualidade em tempo real por meio do *Transcheck*, visualização do texto de chegada com formatação preservada e suporte multilíngue. Essa configuração permite que

tradutores avaliem a adequação do software às suas necessidades profissionais antes de adquirir a licença completa.

O processo de criação do projeto, nomeado *ICD-11-MMS-TM2-en-pt*, no *Wordfast Pro,* segue uma sequência estruturada de configurações. Primeiramente, define-se o projeto com o par linguístico inglês-português. Em seguida, acessa-se a aba *Glossary* e utiliza-se a função *Import* para incorporar o glossário terminológico de referência criado para consulta (marcar *Read Only*), *TB_ICD-11-MMS-en-pt.tbx*, no formato TBX (*Import Report: 36049 terms have been imported*). O próximo passo consiste em abrir a aba *Translation Memory* e selecionar *Create* para criar a memória de tradução específica do projeto, denominada *TM-ICD-11-MMS-TM2-en-pt*. Na aba *Source Files*, utiliza-se a opção *Add Files* para importar o arquivo de texto de partida *ICD-11-MMS-TM2-en.txt* que será traduzido. Após completar todas essas configurações essenciais do glossário, da memória de tradução e do texto de partida, finaliza-se o processo clicando em *Create Project* para abrir efetivamente o ambiente de tradução. Esse procedimento garante que todos os recursos terminológicos e de memória estejam adequadamente integrados antes do início da tradução propriamente dita. A memória criada, *TM_ICD-11-MMS-TM2-en-pt*, será destinada a armazenar as traduções realizadas durante o processo.

Considerando a natureza altamente especializada do conteúdo médico, o processo de tradução é executado sem o auxílio de mecanismos de tradução automatizada, privilegiando-se consultas sistemáticas ao glossário de referência através do sistema de busca integrado da ferramenta. Esse procedimento assegura maior precisão terminológica e consistência conceitual. Para ilustrar a metodologia empregada, na tradução do termo *Abdominal cavity disorders*, inicia-se a busca por *Abdominal cavity* no glossário, que retorna múltiplas ocorrências contextualizadas indicando *Cavidade abdominal* como equivalente terminológico estabelecido. O componente *Disorder*, por sua vez, apresenta diversas opções tradutórias, sendo transtorno a mais frequente, seguida por distúrbio, deficiência e complicação.

Para determinar a escolha mais apropriada ao contexto específico, a frequência de uso e os padrões contextuais podem ser analisados mediante ferramentas de análise linguística como *Antconc* (*Anthony*, 2024a) e *AntPconc* (*Anthony*, 2018), que possibilitam examinar concordâncias e comportamentos terminológicos em corpora especializados da área médica.

Com base nesta análise contextual e terminológica sistemática, *Abdominal cavity disorders* é traduzido como Transtornos da cavidade abdominal, preservando a consistência terminológica estabelecida no glossário de referência da OMS e respeitando os padrões de uso consolidados na literatura médica especializada.

Devido ao caráter específico do Módulo II da ICD-11, referente às medicinas tradicionais (MT2), foi identificada uma lista de 188 palavras que não constam no glossário de referência estabelecido. Entre os exemplos das dez primeiras por ordem alfabética (gerada pelo *Antconc*) encontram-se: *Absorptive, accumulated, Achillodynia, Adenoiditis, affliction, Ageusia, Aggravation, agitated, Aiyam*. Alguns desses termos apresentam familiaridade no contexto médico geral, porém não integram a descrição "formal" das doenças conforme a terminologia médica ocidental padronizada na ICD-11, constituindo vocabulário mais característico das práticas de medicina tradicional.

Outros termos são genuinamente específicos dos sistemas médicos tradicionais, como *Aiyam* (Muthiah, K.; Ganesan, K.; Ponnaiah, M.; Parameswaran, S, 2019), que representa um conceito da medicina tradicional malaia referente a um estado de fraqueza ou debilidade pós-parto experimentado por mulheres. Este termo exemplifica a complexidade terminológica do Módulo II, que busca incorporar conceitos de saúde e doença provenientes de diferentes tradições médicas mundiais, muitos dos quais não possuem equivalentes diretos na nosologia[13] médica ocidental. A ausência desses termos no glossário de referência evidencia a necessidade de recursos terminológicos especializados para a tradução adequada deste módulo específico da classificação internacional.

5.2 Matecat

Conforme apresentado no item 3.2.2, o glossário foi elaborado em formato *Excel* compatível com as exigências do sistema *Matecat* (Fondazione Bruno Kessler *et al.*, 2025). O processo de criação do projeto *ICD-11-MMS-TM2-en-pt* no *Matecat* segue uma sequência estruturada de configurações. Após a seleção do inglês como língua de partida e do português como língua de chegada, acessaram-se as configurações através do ícone *Settings* para definir

[13] NOSOLOGIA. [Medicina] Parte da medicina que estuda e classifica doenças. In: DICIO: dicionário online de português. [S. l.]: 7Graus, [2025]. Disponível em: https://www.dicio.com.br/nosologia/. Acesso em: 15 jun. 2025.

os recursos de apoio à tradução. Na aba *Translation Memory and Glossary*, encontra-se o *MyMemory*, uma extensa memória de tradução colaborativa compartilhada entre todos os usuários do *Matecat*, teve suas sugestões desativadas para permitir que o tradutor se concentre exclusivamente na terminologia médica especializada. Mediante a função *+ New resource*, foi criado o glossário *TB_ICD-11-MMS-en-pt.xlsx*, que recebeu o glossário de referência *TB_ICD-11-MMS-en-pt.xlsx* por meio da funcionalidade *Import Glossary*.

Após o upload dos 36.000 termos no glossário, foi criada a memória de tradução *TM_ICD-11-MMS-TM2-en-pt.tmx*, destinada a armazenar as traduções realizadas durante o processo. Ao final do trabalho de tradução, essas informações serão convertidas do formato TMX para o formato XLSX utilizando a ferramenta *Goldpan*, permitindo a validação dos termos traduzidos por especialistas e sua futura reutilização.

Na aba *Machine Translation*, o mecanismo de tradução automática padrão do *Matecat* é o *ModernMT Lite*, uma versão gratuita do *ModernMT* (Translated, 2025a) que não incorpora os recursos avançados. Esse mecanismo foi desativado pelas mesmas razões apresentadas para o *MyMemory* (Translated, 2025b): a necessidade de manter o foco exclusivo na terminologia médica especializada, evitando interferências de sugestões automatizadas que poderiam comprometer a precisão e consistência terminológica exigida na tradução de conteúdo médico altamente especializado, como o da CID-11.

O processo de tradução operacionaliza-se mediante a consulta sistemática ao glossário de referência, seguindo a metodologia descrita anteriormente para o *Wordfast Pro*. O sistema *Matecat* não apresenta as limitações encontradas no *Wordfast Pro* (restrição de 500 unidades de tradução – UT, na versão de demonstração), possibilitando a implementação gratuita de um processo de consulta automatizado mais robusto. Para tanto, foi criada uma memória de tradução de consulta (função *"*Update*"* desativada, ou seja, o modo de leitura da TM), ***TM_ICD-11-MMS-en-pt***, contendo 36.000 unidades de tradução, gerada pela da conversão do arquivo TBX original para o formato TMX, utilizando a ferramenta *Goldpan* (Logrus Global, 2021). Essa configuração permite acesso direto e automatizado ao conjunto completo de equivalências terminológicas, otimizando o fluxo de trabalho tradutório e garantindo maior consistência na aplicação da terminologia especializada durante todo o processo de tradução do material da CID-11.

O processo revelou-se altamente produtivo e seguro do ponto de vista da seleção criteriosa dos dados oficiais da OMS. Além disso, proporcionou acesso simultâneo às informações terminológicas em duas modalidades complementares: consulta tradicional pelo glossário e consulta inovadora mediante sugestões automatizadas da memória de tradução. Essa configuração dupla garante maior confiabilidade na aplicação dos dados terminológicos oficiais, uma vez que os mesmos dados de referência são disponibilizados tanto por meio da busca manual no glossário quanto através das sugestões automáticas geradas pela memória de tradução, criando um sistema de verificação cruzada que minimiza erros e assegura a consistência terminológica ao longo de todo o processo tradutório.

Embora a integração com CAT tools atenda às necessidades práticas de tradutores, a consulta avançada da estrutura terminológica da CID-11 requer ferramentas específicas de análise linguística multilíngue. Nesse contexto, os concordanciadores monolíngues se apresentam como instrumentos valiosos para a exploração sistemática dos recursos terminológicos multilíngues.

6 Implementação em concordanciador monolíngue: *Antconc*

Embora o *Antconc* (Anthony, 2024a) seja originalmente concebido para análise monolíngue, ele se torna particularmente útil para gerenciar consultas aos glossários multilíngues da CID-11 da OMS por meio de um fluxo de trabalho específico, abrindo-se uma sessão separada do programa para cada idioma estudado (Miroir, 2016). Essa abordagem possibilita comparações entre os recursos terminológicos complexos disponíveis nas diferentes versões linguísticas da classificação internacional, permitindo a validação terminológica e estudos tradutórios contrastivos.

Para que o *Antconc* possa realizar análises morfossintáticas avançadas nos glossários multilíngues da CID-11, torna-se necessário um pré-processamento dos dados terminológicos com ferramentas de PLN (Miroir, 2024a). O processo de enriquecimento, etapa fundamental que antecede a análise no concordanciador, é descrito a seguir.

6.1 Enriquecimento de corpus com o spaCy

O enriquecimento de corpus consiste no processo de adicionar informações linguísticas estruturadas a textos brutos, transformando-os em recursos mais informativos para a análise. Esse processo inclui a adição de metadados, anotações morfossintáticas, semânticas e estruturais que facilitam buscas complexas e análises linguísticas avançadas. O enriquecimento pode envolver a etiquetagem gramatical, a lematização, a análise sintática, a identificação de entidades nomeadas e a marcação de relações semânticas.

O projeto *Universal Dependencies* (Universal Dependencies Consortium, 2024) estabeleceu 17 etiquetas [14] morfossintáticas universais para facilitar análises linguísticas consistentes entre idiomas. Essa padronização (*UPOS Tagging*) permite comparações interlinguísticas precisas e análises multilíngues consistentes, essenciais para pesquisas em terminologia médica internacional.

O spaCy é uma biblioteca de PLN de código aberto para Python, desenvolvida pela *Explosion AI* (Honnibal, Montani, 2024), que oferece modelos pré-treinados de PLN para diversos idiomas, incluindo português, inglês, espanhol e francês. Os modelos variam em tamanho e precisão (*sm: small, md: medium, lg: large*), sendo os maiores mais precisos, mas computacionalmente mais exigentes. Para a análise terminológica da área médica, os modelos "grandes" oferecem melhor desempenho na identificação de termos técnicos e estruturas terminológicas mais complexas.

Para executar esses modelos de uma forma amigável, Laurence Anthony desenvolveu uma nova versão do *TagAnt* (Anthony, 2024b), ferramenta complementar ao *Antconc*, especializada no enriquecimento automático de corpus através de etiquetagem morfossintática. Essa ferramenta processa textos brutos, adicionando informações gramaticais (categoria morfossintática e lema) que facilitam buscas criteriosas no *Antconc*, similares às realizadas no *Sketch Engine* (*Lexical Computing*, 2025). O *TagAnt* oferece suporte a diferentes sistemas de etiquetagem morfossintática e foi preparado para trabalhar com modelos multilíngues externos de PLN da biblioteca *spaCy*. Sua integração metodológica com o *Antconc*

[14] ADJ (Adjetivo), ADP (Adposição), ADV (Advérbio), AUX (Auxiliar), CCONJ (Conjunção coordenativa), DET (Determinante), INTJ (Interjeição), NOUN (Substantivo), NUM (Numeral), PART (Partícula), PRON (Pronome), PROPN (Nome próprio), PUNCT (Pontuação), SCONJ (Conjunção subordinativa), SYM (Símbolo), VERB (Verbo), X (Outros)

possibilita a criação de processos de busca eficientes para análises linguísticas e terminológicas avançadas.

A operacionalização do *TagAnt* para enriquecimento morfossintático dos glossários em inglês e português foi realizada utilizando categorias morfossintáticas, porém sem a inclusão de lemas (MIROIR, 2024a). Esta decisão justifica-se pela natureza do corpus, que apresenta predomínio de estruturas nominais e variabilidade de desinências verbais reduzida, sendo as formas verbais empregadas predominantemente com função adjetival. O processo fundamentou-se nos arquivos previamente preparados e incluiu a escolha do modelo de linguagem (PLN) correspondente à língua de trabalho, a definição do tipo de representação da etiquetagem (*word+pos*) e a renomeação dos arquivos etiquetados com o sufixo *tagged*, conforme ilustrado no Quadro 7 seguinte:

Quadro 6 – Configurações básicas do *TagAnt* para cada modelo de linguagem (*PLN*)[15]

ID	Arquivos preparados	Modelo de linguagem (PLN)	Etiquetagem	Arquivos etiquetados
1	ICD-11-MMS-en-1.txt	English (en_core_web_lg-3.7.0)	word+pos	ICD-11-MMS-en-1-tagged.txt
2	ICD-11-MMS-pt-1.txt	Portuguese (pt_core_news_lg-3.7.0)	word+pos	ICD-11-MMS-pt-1-tagged.txt
3	ICD-11-MMS-fr-1.txt	French (fr_core_news_lg-3.7.0)	word+pos	ICD-11-MMS-fr-1-tagged.txt
4	ICD-11-MMS-es-1.txt	Spanish (es_core_news_lg-3.7.0)	word+pos	ICD-11-MMS-es-1-tagged.txt

Fonte: elaborado pelo autor.

Os arquivos etiquetados gerados com enriquecimento morfossintático estão disponíveis para exploração no *Antconc*. No contexto do processo tradutório, os arquivos são carregados em duas sessões distintas do *Antconc*, conforme a Figura 9 – Processo de Tradução embasada em Corpora (TEsC) (Miroir, 2016, p.22), para executar o projeto de análise em paralelo e observar as estruturas presentes no texto original em inglês (janela A do *Antconc*) e nas respectivas traduções em português (janela B do *Antconc*), eventualmente, em francês (janela C do *Antconc*) ou em espanhol (janela D do *Antconc*).

[15] Os recursos criados nesta pesquisa encontram-se disponíveis no repositório de acesso aberto Figshare (Miroir, 2025)

Para realizar esse tipo de pesquisa avançada, é necessário usar as funcionalidades do *Corpus Manager* do *Antconc* para possibilitar a leitura e a extração dos padrões sintáticos pesquisados. A criação do corpus é executada ao clicar em *"Create"*, para isso, o *Antconc* cria um banco de dados (.db) que pode ser exportado para arquivamento e reuso compartilhado em outras máquinas: *ICD-11-MMS-en-1-tagged.db*.

Para atender aos objetivos propostos nesta pesquisa, a visualização das etiquetas morfossintáticas torna-se um procedimento metodológico relevante. Neste caso, a ativação passa pelos seguintes passos: *Settings > Tools Settings > KWIC > Display Type: Type+POS* (sem lema, chamados "Headwords"). Por exemplo, o padrão pesquisado no *Antconc*, carregado com corpus em português (janela B), *transtorno*_NOUN *_ADP o_DET *_NOUN *_ADJ *_ADJ* da ICD-11 retorna termos compostos complexos que seguem uma estrutura sintática específica na terminologia médica em português:

(1) *transtorno*_NOUN* >>> Substantivo núcleo começando com "transtorno" (permite variações com o plural, como "transtornos");
(2) **_ADP* >>> Preposição (qualquer preposição: de, com, por, etc.);
(3) *o_DET* >>> Artigo definido masculino "o";
(4) **_NOUN* >>> Substantivo complementar (especifica o tipo/área do transtorno);
(5) **_ADJ* >>> Primeiro adjetivo qualificador;
(6) **_ADJ* >>> Segundo adjetivo qualificador.

Os resultados obtidos incluem:

- *Transtornos_NOUN de_ADP o_DET sistema_NOUN hormonal_ADJ hipofisário_ADJ* (janela B) >>> com padrão correspondente em inglês >>> *Disorders_NOUN of_ADP the_DET pituitary_ADJ hormone_NOUN system_NOUN* (janela A);
- *Transtornos_NOUN de_ADP o_DET sistema_NOUN hormonal_ADJ gonadal_ADJ* (janela B) >>> com padrão correspondente em inglês >>> *Disorders_NOUN of_ADP the_DET gonadal_~~PROPN~~* (correto: ADJ) *hormone_NOUN system_NOUN* (janela A).

Para a tradução terminológica médica baseada na CID-11 da OMS, a metodologia proposta revela-se particularmente pertinente, constituindo um complemento essencial às

análises anteriormente desenvolvidas no item 5 . O arquivo com a terminologia da Medicina Tradicional do Módulo II (TM2, em inglês), *ICD-11-MMS-TM2-en.txt*, em fase de tradução, foi etiquetado com o modelo (en_core_web_lg-3.7.0) e renomeado *ICD-11-MMS-TM2-en-tagged.txt.*

A etiquetagem morfossintática permite identificar padrões sintáticos compostos pelos termos que ainda não foram oficialmente traduzidos na CID-11, o que compromete a validação da terminologia através de consultas aos bancos de dados TBX ou XLSX nas CAT Tools. Essa lacuna terminológica representa um obstáculo significativo no processo tradutório, uma vez que a ausência de equivalentes validados limita a manutenção da consistência terminológica. A solução metodológica proposta consiste em identificar padrões sintáticos similares preexistentes no corpus em inglês e estabelecer correspondências estruturais com os padrões equivalentes em português. Essa abordagem permite criar um mapeamento sistemático entre as estruturas sintáticas das duas línguas, facilitando a identificação de lacunas terminológicas e orientando a criação de traduções consistentes com os padrões já estabelecidos na nomenclatura médica internacional.

A tradução de termos compostos, como *Accumulation of kapha pattern (TM2),* mostra a dificuldade de adaptar a CID-11 para diferentes sistemas médicos (Índia, China). Esse exemplo tem dois problemas principais. Primeiro, o termo *kapha* é específico da medicina ayurvédica e designa uma das três energias vitais que controlam o corpo (Círculo Ayurveda, 2024). Esse conceito não existe na medicina ocidental. Segundo, o termo *accumulation* (acumulação) tem sentidos diferentes. Na medicina ocidental, acumulação significa o acúmulo físico de substâncias no corpo, mas não consta no glossário da CID-11. Na medicina tradicional ayurvédica, refere-se a desequilíbrios energéticos que podem não ter sinais físicos visíveis.

Essa particularidade conceitual não se restringe ao Módulo II da ICD-11, pois o Módulo I (TM1), dedicado à Medicina Tradicional Chinesa, apresenta estruturas sintáticas similares que evidenciam a sistematicidade desses padrões terminológicos: (1) *X accumulation pattern (TM1)*, exemplificado por Bladder heat accumulation pattern (TM1); e (2) *X accumulation in Y pattern (TM1)*, como em *Blood accumulation in the bladder pattern (TM1)*. Essa recorrência estrutural sugere que o conceito de *accumulation* representa um denominador comum nos sistemas médicos tradicionais. No entanto, observam-se duas concepções distintas: os padrões

energéticos (como em *heat accumulation pattern*) e o acúmulo físico de substâncias (como em *blood accumulation in*).

A tradução adequada desses termos exige, portanto, não apenas competência linguística, mas uma compreensão aprofundada dos paradigmas epistemológicos subjacentes a cada sistema médico, conforme apresentado no Quadro 8 seguinte:

Quadro 7 – Trechos etiquetados da ICD-11 (TM1) e a sugestão de tradução para CID-11 (TM2)

ICD-11 (TM1)	CID-11 (TM1) - Tradução oficial da OMS
Bladder_NOUN heat_NOUN accumulation_NOUN pattern_NOUN	Padrão_NOUN de_ADP Acúmulo_~~PROPN~~ de_ADP Calor_~~PROPN~~ em_ADP a_DET Bexiga_~~PROPN~~
Blood_NOUN accumulation_NOUN in_ADP the_DET bladder_NOUN pattern_NOUN	Padrão_NOUN de_ADP Acúmulo_~~PROPN~~ de_ADP Sangue_~~PROPN~~ em_ADP a_DET Bexiga_~~PROPN~~
ICD-11 (TM2)	**CID-11 (TM2) - Sugestão de tradução para OMS**
Accumulation_NOUN of_ADP Vata_**PROPN** pattern_NOUN	Padrão_NOUN de_ADP acúmulo_NOUN de_ADP Vata_~~PROPN~~
Accumulation_NOUN of_ADP pitta_NOUN pattern_NOUN	Padrão_NOUN de_ADP acúmulo_NOUN de_ADP pita_NOUN
Accumulation_NOUN of_ADP kapha_**PROPN** pattern_NOUN	Padrão_NOUN de_ADP acúmulo_NOUN de_ADP kapha_NOUN

Fonte: elaborado pelo autor.

Observa-se que tanto no *ICD-11-MMS-en-1-tagged.txt* quanto nos arquivos *ICD-11-MMS-pt-1-tagged.txt e ICD-11-MMS-TM2-en-1-tagged.txt* aparecem etiquetas como PROPN (destacadas em negrito e riscadas no quadro acima) que identificam os nomes próprios. Essa ocorrência se deve ao fato de os modelos utilizados terem sido treinados em corpora generalistas: o en_core_web_lg-3.7.0 (Explosion AI, 2023a) foi treinado em textos da web, enquanto o pt_core_news_lg-3.7.0 (Explosion AI, 2023b) foi treinado em textos jornalísticos de plataformas de agregação automatizada de notícias, como Google News, Yahoo News e outras fontes similares.[16] Ambos os modelos de linguagem apresentam limitações quando aplicados à terminologia médica especializada, uma vez que foram desenvolvidos distantes da especificidade terminológica do domínio da CID-11. Além disso, esses modelos são sensíveis à capitalização das palavras, tendendo a classificar termos iniciados com maiúscula como nomes

[16] De modo similar, os modelos *fr_core_news_lg* (EXPLOSION AI, 2023c) e *es_core_news_lg* (EXPLOSION AI, 2023d) foram treinados com corpus de notícias em francês e espanhol, respectivamente.

próprios (PROPN), mesmo quando se trata de termos técnicos especializados da nomenclatura médica, conforme descrito no Quadro 9 seguinte:

Quadro 8 – Influência da capitalização das palavras para identificações como PROPN

1) Termo em inglês
Spleen deficiency with dampness accumulation pattern (TM1)
Spleen_**PROPN** deficiency_NOUN with_ADP dampness_NOUN accumulation_NOUN pattern_NOUN
2) Termo traduzido em português, conforme a capitalização original no arquivo da OMS
Padrão de Deficiência do Baço com acúmulo de Umidade (MT1)
Padrão_NOUN de_ADP Deficiência_~~**PROPN**~~ do_ADP Baço_~~**PROPN**~~ com_ADP acúmulo_NOUN de_ADP Umidade_NOUN
3) Termo traduzido em português, conforme a capitalização do português padrão
Padrão de deficiência do baço com acúmulo de umidade
Padrão_NOUN de_ADP deficiência_NOUN do_ADP baço_NOUN com_ADP acúmulo_NOUN de_ADP umidade_NOUN

Fonte: elaborado pelo autor.

No exemplo 1 “Termo em inglês”, o modelo *en_core_web_lg-3.7.0* conseguiu etiquetar corretamente a informação completa, porém classificou erroneamente a primeira palavra capitalizada como nome próprio (PROPN), o que não ocorreu com a palavra *“*Padrão*”* nos exemplos 2 e 3. No exemplo 2, o modelo ***pt_core_news_lg-3.7.0*** classificou incorretamente duas palavras capitalizadas, *“*Deficiência*”* e *“*Baço*”*, como nomes próprios, embora tenha etiquetado corretamente *“*Umidade_NOUN*”*. No exemplo 3, o modelo apresentou desempenho adequado após a formatação correta dos substantivos comuns, utilizando apenas a inicial maiúscula, conforme as convenções ortográficas padrão.

Esses resultados evidenciam que ambos os modelos, treinados em corpora generalistas, distantes da especificidade terminológica médica, apresentam sensibilidade excessiva à capitalização, interpretando termos técnicos especializados como nomes próprios quando apresentados com formatação não convencional. Essa limitação compromete a precisão da etiquetagem morfossintática em textos médicos especializados, sugerindo a necessidade de modelos específicos para o domínio médico ou de pré-processamento adequado dos textos antes da análise.

O modelo spaCy *en_core_web_lg-3.7.0* foi testado no *TagAnt* com o padrão “word+pos_tag*”* do sistema de etiquetagem (pos_tag) do *Penn Treebank* (Marcus, M. P.;

Marcinkiewicz, M. A.; Santorini, B., 1993), não disponível para *pt_core_news_lg-3.7.0*. No entanto, os resultados de etiquetagem de nomes próprios (PROPN e NNP) são similares ao "word+pos", ou seja, respectivamente, 180 oc. NNP e 181 oc. PROPN, identificando exatamente as mesmas palavras (

Quadro 9). Os modelos *spaCy* da versão 3.8.0 disponíveis *on-line* foram testados por meio de um *script* em *Python*, uma vez que o *TagAnt* utiliza apenas a versão 3.7.0. Os testes demonstraram melhorias na etiquetagem PROPN em termos de acurácia: 181 oc. PROPN (3.7.0) e 114 oc. PROPN (3.8.0). No entanto, diante dessa limitação, o modelo de PLN *Stanza* poderia representar uma alternativa mais eficaz para essa tarefa específica.

6.2 Enriquecimento de corpus com Stanza

O *Stanza* é uma biblioteca de PLN desenvolvida pela Universidade de Stanford que oferece um *pipeline* neural completamente integrado para análise linguística multilíngue em 66 idiomas. Desenvolvido por Qi *et al.* (2020), a ferramenta implementa uma arquitetura neural unificada que funciona para qualquer desses idiomas, executando tarefas como a tokenização, a segmentação de sentenças, a etiquetagem morfossintática (*POS tagging*) com características morfológicas universais, a lematização, a análise sintática (*dependency parsing*) utilizando redes neurais e o reconhecimento de entidades nomeadas (*NER*) com representações contextualizadas. Diferentemente do *spaCy*, que prioriza velocidade e eficiência para aplicações industriais, o *Stanza*, mais lento, foca a precisão linguística para pesquisa acadêmica, sendo treinado em centenas de *datasets*, incluindo os *Universal Dependencies treebanks*. Segundo QI *et al.* (2020), a biblioteca *Stanza* demonstra performances superiores em análise sintática, alcançando 92,49% de acurácia em *UPOS tagging*[17], comparado aos resultados considerados inferiores do *spaCy* nos mesmos testes padronizados. O *Stanza* representa uma evolução metodológica em relação ao *spaCy*, oferecendo maior precisão para anotação automática de grandes corpora multilíngues através do padrão *Universal Dependencies*,

[17] UPOS Tagging: Sistema de etiquetagem gramatical do projeto Universal Dependencies (UD), que utiliza um conjunto universal e simplificado de categorias (ex.: VERB, NOUN). Diferencia-se de padrões como o do Penn Treebank por priorizar a comparabilidade entre línguas, enquanto o último oferece um conjunto de etiquetas mais extenso e específico para a sintaxe de cada idioma.

características essenciais para pesquisa em Linguística de Corpus quantitativa que demanda análise sintática profunda e consistência interlinguística.

Para executar o *Stanza*, é aconselhável criar um ambiente virtual Python denominado "stanza" onde serão instaladas as bibliotecas necessárias para a execução dos scripts de processamento de dados. O *Stanza* pode ser baixado diretamente de seu repositório no GitHub ou através do PyPI, utilizando o comando *pip install stanza*. O script em Python, desenvolvido com auxílio do assistente de inteligência artificial *Copilot* (Microsoft, 2025), permite gerar a etiquetagem morfossintática do material de forma similar à produzida pelo *spaCy*, no formato *"*palavra_POS*"*, conforme exemplificado em *"*doenças_NOUN*"*. De fato, o *Stanza* apresenta velocidade de processamento inferior em relação ao *spaCy*. Para etiquetar o arquivo em análise *ICD-11-MMS-TM2-en.txt* (733 termos médicos compostos), foram necessários um minuto e 33 segundos de processamento [18] com o *Stanza*, contrastando com menos de 8 segundos utilizando o *spaCy* executado no *TagAnt*. O resultado foi gerado no arquivo: *ICD-11-MMS-TM2-en-tagged-py-stanza.txt*

7 Análises de dados e comparação dos dois modelos de PLN

Devido aos problemas de etiquetagem inconsistente da categoria PROPN (nomes próprios), a análise comparativa revela que tanto o *spaCy* quanto o *Stanza* apresentam limitações substanciais na classificação de nomes próprios (PROPN) em terminologia médica da CID-11, com taxas de erro superiores a 95% em ambos os casos. Esse resultado evidencia um problema estrutural fundamental: modelos de PLN treinados em corpora generalistas (textos da web ou de notícias) enfrentam desafios significativos quando aplicados a domínios altamente especializados, como a terminologia médica internacional CID-11. Ambos os sistemas demonstram a mesma confusão conceitual entre especificidade semântica (precisão técnica elevada dos termos médicos) e propriedade nominal (características de nomes próprios), aplicando erroneamente a lógica de que termos desconhecidos, com maiúscula inicial, devem ser categorizados como nomes próprios.

[18] Máquina usada para execução do processamento: (1) Processador Intel(R) Core(TM) i7 2.80GHz (2.80 GHz) ; (2) Memória RAM instalada de 16,0 GB.

7.1 Limitações dos modelos de PLN generalistas em terminologia médica especializada

A divergência entre o contexto de treinamento generalista e a aplicação médica especializada compromete fundamentalmente a confiabilidade de ambos os modelos para projetos críticos, como a tradução e a análise morfossintática da CID-11, exigindo um desenvolvimento de soluções específicas para o domínio médico ou implementação de protocolos de validação manual para garantir precisão terminológica em contextos clínicos. Para analisar essa limitação estrutural, realizou-se uma análise preliminar em quatro etapas:

1) Levantamento quantitativo das palavras etiquetadas como PROPN com o *Antconc*

Quadro 9 - Levantamento quantitativo das palavras etiquetadas como PROPN com o *Antconc*[19]

Modelo	Arquivo usado	Total oc.	Duplicatas	Total oc. únicas	Total oc. sem maiúsculas
SpaCy 3.7.0	ICD-11-MMS-TM2-en-tagged-ta-*spaCy*.txt	181	60	121	6
Stanza	ICD-11-MMS-TM2-en-tagged-py-stanza.txt	91	24	67	0

Fonte: elaborado pelo autor.

O Quadro 10 demonstra que ambos os sistemas classificam incorretamente os termos médicos como nomes próprios. O spaCy registra o dobro de ocorrências do *Stanza* (181 versus 91 e 121 versus 67 nas ocorrências únicas), evidenciando a dimensão do problema. Um dado particularmente revelador são as seis ocorrências sem maiúsculas etiquetadas como PROPN pelo spaCy, o que aponta para uma falha que vai além da simples análise ortográfica. Esses resultados confirmam que os modelos não são adequados para processar a terminologia especializada da CID-11 sem ajustes específicos.

[19] Os recursos criados nesta pesquisa encontram-se disponíveis no repositório de acesso aberto Figshare (Miroir, 2025)

2) Classificações corretas identificadas por *spaCy* e *Stanza* (dois acertos para ambos):
    - Bartholin_PROPN (epônimo médico - glândula de Bartholin)
    - Ludwig_PROPN (epônimo médico - angina de Ludwig)
3) Análise detalhada das classificações incorretas: *spaCy* (119 oc.) e *Stanza* (65 oc.)
    - Termos médicos básicos que deveriam ser NOUN: *Reproductive, System, Spleen, Bone etc.*
    - Condições clínicas que deveriam ser NOUN: *Diarrhoea, Eczema, Gout, Hernia, Jaundice etc,*
    - Adjetivos que deveriam ser ADJ: *Benign, Deep, Hereditary, Inflammatory etc.*
    - Termos ayurvédicos que deveriam ser NOUN: *Vata, Pitta, Kapha, Ojas etc.*
4) Análise da capitalização: A capacidade de respeitar convenções de capitalização é fundamental para PROPN, pois o *spaCy* classifica palavras em minúsculas (*ano, bulbi, dosha, neonatarum, thrombophilia, vāta*) como PROPN, desconsiderando a regra básica de que nomes próprios iniciam com maiúscula.

Quanto à questão da generalização, existe uma divergência natural entre o treinamento original dos modelos de linguagem de PLN (web/notícias) e o domínio médico. Embora o corpus CID-11 seja altamente especializado, essa divergência pode afetar a extrapolação dos resultados. A escolha metodológica de focar em PROPN foi acertada, pois é precisamente nessa categoria que o extenso conhecimento adquirido pelos grandes modelos de linguagem (LLM) se torna fundamental para distinguir adequadamente entre terminologia técnica especializada e nomes próprios genuínos no contexto médico.

7.2 Propostas para aprimoramento dos modelos de etiquetagem morfossintática

A recente literatura científica confirma as limitações observadas neste estudo sobre modelos de processamento de linguagem natural generalistas em contextos médicos. Com efeito, conforme Saeed e Naveed (2022), os sistemas de PLN treinados em corpora generalistas apresentam inadequação substancial quando aplicados à terminologia médica, evidenciando a necessidade de adaptação específica para domínios especializados. Esses achados de Saeed e Naveed corroboram os resultados da aplicação do projeto Tradução Médica em Contexto

(TraMeC), descrita neste artigo, onde *spaCy* e *Stanza* demonstraram taxa de erro superior a 95% na classificação de nomes próprios (PROPN) em terminologia da CID-11. A convergência entre as observações desta pesquisa e a literatura estabelecida valida a abordagem metodológica de combinar recursos oficiais da OMS com tecnologias de PLN adaptadas para superar essas deficiências.

Para melhorar a identificação de PROPN no *spaCy*, especialmente para a terminologia médica como a da CID-11, sugere-se uma abordagem híbrida que combine diferentes estratégias de processamento, organizadas em ordem crescente de complexidade técnica.

A primeira abordagem consiste na implementação de pós-processamento baseado em regras utilizando expressões regulares, representando a estratégia mais direta e acessível (Saeed; Naveed, 2022). Esse componente aproveitaria padrões linguísticos característicos da terminologia médica, como sufixos latinos (-osis, -itis para inflamações, -oma para tumores, -emia para condições sanguíneas) e contexto sintático específico, para aplicar correções imediatas nos resultados do modelo.

A segunda abordagem, mais técnica e recomendável, consiste no *fine-tuning* direcionado do modelo pré-treinado utilizando dados específicos do domínio médico. Essa metodologia, baseada nas diretrizes do *spaCy* (Explosion AI, 2025), focalizaria a adaptação do *tagger* que preserva o conhecimento linguístico geral enquanto aprimora a classificação morfossintática em terminologia médica. O procedimento envolveria a preparação de um corpus anotado derivado dos glossários CID-11, configuração de arquivo de treinamento específico e execução do *fine-tuning.*[20] Essa abordagem representa uma solução escalável e tecnicamente robusta, permitindo correção das limitações identificadas na classificação PROPN e estabelecendo um modelo adaptado aplicável consistentemente a novos textos médicos.

A terceira abordagem consiste na avaliação sistemática de modelos de PLN especificamente desenvolvidos para o domínio médico. Essa abordagem avaliaria o desempenho de sistemas especializados, como *BioBERT* (Alsentzer, 2019), *ClinicalBERT* (Peng, Y.; Yan, S.; Lu, Z., 2019), bem como modelos biomédicos específicos do *spaCy*, *scispaCy* (Neumann, M.; King, D.; Beltagy, I.; Ammar, W, 2019) e *Stanza* "*biomedical models*" (Zhang, 2021), que foram treinados exclusivamente com corpora clínicos, potencialmente oferecendo

[20] Quick Start. Disponível em: https://spacy.io/usage/training/#quickstart . Acesso em: 25 jun. 2025.

maior precisão na classificação de terminologia médica da CID-11. Essa estratégia representa uma evolução natural do projeto TraMeC, estabelecendo um novo paradigma que combine dados oficiais da OMS com modelos de PLN otimizados para contextos biomédicos.

8 Conclusão

A aplicação do projeto Tradução Médica em Contexto (TraMeC), descrita neste artigo, demonstrou a viabilidade de criar ferramentas terminológicas de alta precisão através da combinação entre dados oficiais da OMS (CID-11) e tecnologias de PLN. A metodologia desenvolvida resultou em 36.049 pares terminológicos validados em quatro idiomas, convertidos em glossários TBX e XLSX compatíveis com as principais *CAT Tools*. A implementação prática confirmou a funcionalidade dos recursos em ferramentas como *Wordfast Pro* e *Matecat*, enquanto a validação cruzada por meio de concordanciadores fortaleceu a confiabilidade dos alinhamentos terminológicos. Contudo, a pesquisa revelou limitações significativas na análise morfossintática automatizada, com taxa de erro superior a 95% na classificação de nomes próprios, evidenciando que modelos treinados em corpora generalistas enfrentam dificuldades com terminologia médica altamente especializada. Os recursos desenvolvidos representam um avanço concreto para a padronização terminológica em tradução médica, democratizando o acesso à terminologia oficial através de formatos profissionais. A disponibilização no *Figshare* (Miroir, 2025) assegura sustentabilidade e adaptabilidade pela comunidade usuária.

Para desenvolvimentos futuros, recomenda-se o uso de modelos PLN específicos ajustados (*fine tuned*) para o domínio médico, a expansão para outros idiomas da CID-11 e pesquisas sobre aplicações a outras nomenclaturas médicas internacionais (SNOMED CT). Essa aplicação do projeto TraMeC confirma o papel da LC como ponte entre teoria linguística e aplicações práticas, estabelecendo um modelo replicável para outras áreas que demandem precisão terminológica e contribuindo significativamente tanto para os estudos tradutológicos quanto para a prática da tradução médica especializada.

Referências

ALSENTZER, E. et al. **Publicly available clinical BERT embeddings**. arXiv preprint, [s. l.], arXiv:1901.08746, 2019. Disponível em: https://arxiv.org/pdf/1901.08746. Acesso em: 25 jun. 2025.

ANTHONY, L. **AntPconc**. Versão 1.2.1. Tokyo: Waseda University, 2018. Software. Disponível em: https://www.laurenceanthony.net/software/antpconc/. Acesso em: 20 maio 2025.

ANTHONY, L. ***Antconc***. Versão 4.3.1. Tokyo: Waseda University, 2024a. Software. Disponível em: https://www.laurenceanthony.net/software/*Antconc*/. Acesso em: 20 maio 2025.

ANTHONY, L. **TagAnt**: Windows installer. Versão 2.1.1. Tokyo: Waseda University, 2024b. Software de etiquetagem morfossintática. Disponível em: https://www.laurenceanthony.net/software/tagant/. Acesso em: 20 maio 2025.

CHAMPOLLION INC. **Wordfast Pro**. Versão 9.12.0. [S. l.]: Champollion Inc., jan. 2025. Software. Disponível em: https://www.wordfast.com/products/wordfast_pro#. Acesso em: 2 maio 2025.

CÍRCULO AYURVEDA. **Os três doshas ayurvédicos**: Vata, Pitta e Kapha explicados. [S. l.], [2024]. Disponível em: https://circuloayurveda.com.br/os-tres-doshas-ayurvedicos-vata-pitta-e-kapha-explicados/ . Acesso em: 15 jun. 2025.

EXPLOSION AI. **Training pipelines & models**. In: SPACY: industrial-strength natural language processing. Berlin: Explosion AI, 2025. Documentação técnica. Disponível em: https://spacy.io/usage/training/. Acesso em: 17 jun. 2025.

EXPLOSION AI. **en_core_web_lg**: English language model for spaCy. Versão 3.7.0. Berlin: Explosion AI, jul. 2023a. Modelo de processamento de linguagem natural. Disponível em: https://spacy.io/models/en. Acesso em: 2 jun. 2025.

EXPLOSION AI. **pt_core_news_lg**: Portuguese language model for spaCy. Versão 3.7.0. Berlin: Explosion AI, jul. 2023b. Modelo de processamento de linguagem natural. Disponível em: https://spacy.io/models/pt. Acesso em: 2 jun. 2025.

EXPLOSION AI. **fr_core_news_lg**: French language model for spaCy. Versão 3.7.0. Berlin: Explosion AI, jul. 2023c. Modelo de processamento de linguagem natural. Disponível em: https://spacy.io/models/fr. Acesso em: 2 jun. 2025.

EXPLOSION AI. **es_core_news_lg**: Spanish language model for spaCy. Versão 3.7.0. Berlin: Explosion AI, jul. 2023d. Modelo de processamento de linguagem natural. Disponível em: https://spacy.io/models/es. Acesso em: 2 jun. 2025.

FERNÁNDEZ-PARRA, M. **Terminology Management for Translators**: a guide for students, trainers and professionals. 1. ed. London: Routledge, 2025. (Routledge Introductions to Translation and Interpreting). Disponível em: https://api.pageplace.de/preview/DT0400.9781040351437_A50830233/preview-9781040351437_A50830233.pdf. Acesso em: 18 out. 2025.

FONDAZIONE BRUNO KESSLER; TRANSLATED SRL; UNIVERSITÉ DU MAINE; UNIVERSITY OF EDINBURGH. **Matecat**: computer-assisted translation tool. [S. l.]: FBK, 2025. Software colaborativo de código aberto. Disponível em: https://www.matecat.com. Acesso em: 15 jun. 2025.

FUNG, K. W. *et al*. **Promoting interoperability between SNOMED CT and ICD-11**: lessons learned from the pilot project mapping between SNOMED CT and the ICD-11 Foundation. Journal of the American Medical Informatics Association, v. 31, n. 8, p. 1631-1637, Aug. 2024. DOI: https://doi.org/10.1093/jamia/ocae143.

HARRISON, J. E.; WEBER, S.; JAKOB, R.; CHUTE, C. G. **ICD-11**: an international classification of diseases for the twenty-first century. BMC Medical Informatics and Decision Making, [s. l.], v. 21, n. 206, 2021. DOI: https://doi.org/10.1186/s12911-021-01534-6.

HONNIBAL, M.; MONTANI, I. **spaCy**: Industrial-strength Natural Language Processing in Python. [S. l.]: Explosion AI, 2024. Disponível em: https://spacy.io . Acesso em: 25 jun. 2025.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. **ISO 30042:** systems to manage terminology, knowledge and content — TermBase eXchange (TBX). Geneva: ISO, 2019a. Disponível em: https://www.iso.org/standard/62510.html. Acesso em: 15 jun. 2025.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. **ISO/TC 37**: Technical Committee 37: terminology and other language and content resources. Geneva: ISO, 2019b. Disponível em: https://www.iso.org/committee/48104.html. Acesso em: 15 jun. 2025.

KILGRAY TRANSLATION TECHNOLOGIES. **memoQ**: translation management system. Budapeste, c2025. Disponível em: https://www.memoq.com/. Acesso em: 18 out. 2025.

LEXICAL COMPUTING. **Sketch Engine**: corpus analysis software. Brno: Lexical Computing, 2025. Plataforma web para análise linguística de corpus. Disponível em: https://www.sketchengine.eu/ . Acesso em: 15 jun. 2025.

LOGRUS GLOBAL. **Goldpan**: um editor de arquivos TMX/TBX multifuncional e conversor de formatos de arquivos. Versão 3.6.7. Perkasie, PA (EUA), 2021. Desenvolvido por Andrew Kopylev. Disponível em: https://cloud.logrusglobal.com/#*Goldpan*. Acesso em: 10 jun. 2025.

MARCUS, M. P.; MARCINKIEWICZ, M. A.; SANTORINI, B. **Building a large annotated corpus of English**: the *Penn Treebank*. Computational Linguistics, Cambridge, v. 19, n. 2, p. 313-330, 1993. Disponível em: https://aclanthology.org/J93-2004.pdf . Acesso em: 15 maio 2025.

MICROSOFT. **Microsoft Copilot**. Redmond: Microsoft, [2024]. Assistente de inteligência artificial. Disponível em: https://copilot.microsoft.com. Acesso em: 8 jun. 2025.

MIROIR, J.-C. **Search Engine Optimisation (SEO) as a tool for the automated collection of documents for developing the corpora**. ResearchGate, [s. l.], 2016. DOI: 10.5151/sosci-viiieblc-xiii-elc-08_artigo_05.

MIROIR, J.-C. **Processamento de linguagem natural multilíngue com spaCy e análises avançadas de corpora anotados com *Antconc* versão 4**. ResearchGate, [s. l.], 2024a. DOI: 10.13140/RG.2.2.24082.67520 .

MIROIR, J.-C. **Glossários médicos multilíngues para *CAT Tools***: desenvolvimento baseado na CID-11 da OMS [DATASET], Figshare Datacite, 2025. Disponível em: https://figshare.com/projects/TraMeC. Acesso em: 1 outubro 2025.

MUTHIAH, K.; GANESAN, K.; PONNAIAH, M.; PARAMESWARAN, S. **Concepts of body constitution in traditional Siddha texts**: a literature review. Journal of Ayurveda and Integrative Medicine, [s. l.], v. 10, n. 2, p. 131-134, May 2019. DOI: https://doi.org/10.1016/j.jaim.2019.04.002

NEUMANN, M.; KING, D.; BELTAGY, I.; AMMAR, W. **ScispaCy**: fast and robust models for biomedical natural language processing. arXiv preprint, [s. l.], arXiv:1902.07669, 2019. Disponível em: https://arxiv.org/abs/1902.07669 . Acesso em: 25 jun. 2025.

ORGANIZAÇÃO MUNDIAL DA SAÚDE. **Ficha informativa da CID-11**: Classificação Internacional de Doenças 11ª Revisão. Geneva: OMS, 2022. Documento técnico: Ficha informativa da CID-11. Disponível em: https://icd.who.int/pt/docs/ICD-11-factsheet_PT_OMS.pdf. Acesso em: 30 maio 2025.

ORGANIZAÇÃO MUNDIAL DA SAÚDE. Classificação Internacional de Doenças, 11ª Revisão (CID-11): **O padrão global para informações de diagnóstico de saúde**. Genebra, 2025a. Disponível em: https://icd.who.int/pt. Acesso em: 30 maio 2025.

ORGANIZAÇÃO MUNDIAL DA SAÚDE. Classificação Internacional de Doenças, 11ª Revisão: **Navegador CID-11 para Estatísticas de Mortalidade e de Morbidade**. Genebra, 2025b. Disponível em: https://icd.who.int/pt. Acesso em: 30 maio 2025.

ORGANIZAÇÃO MUNDIAL DA SAÚDE. Classificação Internacional de Doenças, 11ª Revisão: **Planilha Excel SimpleTabulation-ICD-11-MMS-pt.xlsx.** Genebra, 2025c. Disponível em: https://icdcdn.who.int/static/releasefiles/2025-01/SimpleTabulation-ICD-11-MMS-pt.zip. Acesso em: 30 maio 2025.

PENG, Y.; YAN, S.; LU, Z. **Transfer learning in biomedical natural language processing**: an evaluation of BERT and ELMo on ten benchmarking datasets. arXiv preprint, [s. l.], Apr. 2019. Disponível em: https://arxiv.org/abs/1906.05474. Acesso em: 25 jun. 2025.

PRIETO RAMOS, F. Ensuring Consistency and Accuracy of Legal Terms in Institutional Translation: The Role of Terminological Resources in International Organizations. In: PRIETO RAMOS, F. (Ed.). **Institutional Translation and Interpreting: Assessing Practices and Managing for Quality**. New York: Routledge, 2020. p. 128-149. DOI: https://doi.org/10.4324/9780429264894-10

QI, P. *et al.* ***Stanza***: a Python natural language processing toolkit for many human languages. In: In Association for Computational Linguistics (ACL) System Demonstrations., 2020, [s. l.]. DOI: https://doi.org/10.48550/arXiv.2003.07082.

RWS TRADOS. **Trados Studio**: software de tradução profissional. Maidenhead: RWS Group, 2025. Disponível em: https://www.rws.com/translation/trados/. Acesso em: 18 out. 2025.

SAEED, N.; NAVEED, H. **Medical terminology-based computing system**: a lightweight post-processing solution for out-of-vocabulary multi-word terms. Frontiers in Molecular Biosciences, [s. l.], v. 9, p. 928530, Aug. 2022. DOI: https://doi.org/10.3389/fmolb.2022.928530

SCHOENING, S. **CAT Tools**: Unlocking the Potential of Computer-Assisted Translation for Global Growth. Phrase Blog, 19 fev. 2024. Disponível em: https://phrase.com/blog/posts/cat-tools/. Acesso em: 18 out. 2025.

SNOMED. **What is SNOMED CT?**. 2025. Disponível em: https://www.snomed.org/what-is-snomed-ct . Acesso em: 15 jun. 2025.

STEFANIAK, K. Terminology work in the European Commission: Ensuring high-quality translation in a multilingual environment. In: SVOBODA, T.; BIEL, Ł.; ŁOBODA, K. (Ed.). **Quality aspects in institutional translation**. Berlin: Language Science Press, 2017. p. 109-121. Disponível em: https://zenodo.org/records/1048192. Acesso em: 18 out. 2025.

TRANSLATED. **ModernMT**: adaptive machine translation platform. [S. l.]: Translated, 2025a. Disponível em: https://www.modernmt.com/ . Acesso em: 10 jun. 2025.

TRANSLATED. **MyMemory**. [S. l.]: Translated, 2025b. Sistema de memória de tradução colaborativa *on-line*. Disponível em: https://mymemory.translated.net/. Acesso em: 10 jun. 2025.

UNIVERSAL DEPENDENCIES CONSORTIUM. **Universal Dependencies**: universal morphosyntactic annotation framework. Stanford: Stanford University, 2024. Framework colaborativo para anotação morfossintática multilíngue. Disponível em: https://universaldependencies.org/. Acesso em: 11 jun. 2025.

ZHANG, Y. *et al.* **Biomedical and clinical English model packages for the *Stanza* Python NLP library**. Journal of the American Medical Informatics Association, [s. l.], v. 28, n. 6, p. 1892-1899, June 2021. DOI: https://doi.org/10.1093/jamia/ocab090.

Recebido em: 03.10.2025 Aprovado em: 02.02.2026