CADERNOS de TRADUÇÃO

Bilinguismo jurídico chinês-português em Macau: análise e alinhamento de corpus com IA

Chinese-Portuguese legal bilingualism in Macao: AI-powered corpus analysis and alignment

Jean-Claude Miroir
Universidade de Brasília
Brasília, Distrito Federal, Brasil
jcmiroir@unb.br
https://orcid.org/0000-0002-5875-9074

Resumo: O artigo explora o bilinguismo jurídico em Macau, uma região administrativa especial da China que adota o princípio "um país, dois sistemas". A análise desenvolvida se concentra na coexistência do chinês e do português na legislação local, destacando a hierarquia linguística e os desafios de tradução jurídica. Com base em uma metodologia robusta, o estudo compila e anota corpora paralelos de textos jurídicos em português e chinês, empregando ferramentas avançadas de inteligência artificial como Bertalign para alinhamento automatizado e spaCy para anotações linguísticas. Devido aos sistemas de escrita distintos entre as línguas, o alinhamento exigiu soluções baseadas em inteligência artificial para superar limitações dos alinhadores convencionais desenvolvidos para línguas românicas. A análise destaca a influência significativa do português na terminologia jurídica em chinês, com uma alta prevalência de *calques* linguísticos. O estudo também aborda os desafios de tokenização da língua chinesa e da tradução automática, propondo soluções práticas e comparando o desempenho de diferentes ferramentas de tradução, como LLMs (modelos de linguagem) e TANs (sistemas de tradução automática neural). A exploração detalhada dos corpora, incluindo análises de n-grams e padrões sintáticos, oferece insights valiosos para a linguística de corpus e a tradução jurídica. As ferramentas computacionais do Sketch Engine possibilitam, através de exemplos práticos extraídos deste corpus alinhado, o estudo dos aspectos específicos da tradução jurídica macauense. O artigo contribui significativamente para o campo da linguística de corpus, especialmente em contextos jurídicos multilíngues, e oferece recursos metodológicos e analíticos valiosos para pesquisadores e profissionais da tradução jurídica.

Palavras-chave: bilinguismo jurídico; Macau; tradução jurídica; linguística de corpus; *calques* linguísticos.

Cadernos de Tradução, 45(Número Especial 3), 2025, e108423
Programa de Pós-Graduação em Estudos da Tradução
Universidade Federal de Santa Catarina, Brasil. ISSN 2175-7968
DOI https://doi.org/10.5007/2175-7968.2025.e108423

Abstract: This article explores legal bilingualism in Macao, a Special Administrative Region of China that adopts the "one country, two systems" principle. The analysis focuses on the coexistence of Chinese and Portuguese in local legislation, highlighting linguistic hierarchy and legal translation challenges. Based on a robust methodology, the study compiles and annotates parallel corpora of legal texts in Portuguese and Chinese, employing advanced artificial intelligence tools such as Bertalign for automated alignment and spaCy for linguistic annotations. Due to the distinct writing systems between the languages, alignment required AI-based solutions to overcome limitations of conventional aligners developed for Romance languages. The analysis highlights the significant influence of Portuguese on Chinese legal terminology, with a high prevalence of linguistic *calques*. The study also addresses the challenges of Chinese language tokenization and machine translation, proposing practical solutions and comparing the performance of different translation tools, such as LLMs (language models) and NMTs (neural machine translation systems). The detailed exploration of the corpora, including n-gram analyses and syntactic patterns, offers valuable insights for corpus linguistics and legal translation. The computational tools of Sketch Engine enable, through practical examples extracted from this aligned corpus, the study of specific aspects of Macanese legal translation. The article contributes significantly to the field of corpus linguistics, especially in multilingual legal contexts, and offers valuable methodological and analytical resources for researchers and legal translation professionals.
Keywords: legal bilingualism; Macao; legal translation; corpus linguistics; linguistic *calques*.

1. Introdução

O Governo da Região Administrativa Especial de Macau (RAEM), estabelecido em 1999 sob o princípio "um país, dois sistemas", foi consagrado na Declaração Conjunta Sino-Portuguesa de 1987 (acordo internacional que definiu a transição de Macau para a China) e na Lei Básica de Macau (constituição local aprovada pela Assembleia Nacional Popular da China em 1993). A RAEM é uma entidade administrativa da República Popular da China (RPC) que goza de alto grau de autonomia, exceto em matéria de defesa e relações exteriores, competências reservadas ao governo central em Pequim. Sua estrutura inclui o Chefe do Executivo, nomeado por Pequim após uma eleição indireta local; o Conselho Executivo, órgão consultivo; a Assembleia Legislativa, unicameral e responsável por funções legislativas; e tribunais independentes, cujo sistema jurídico segue a tradição romano-germânica (herança da administração portuguesa até 1999). Esse sistema contrasta com o da RPC, baseado no direito socialista de inspiração soviética, consolidado após a fundação da RPC em 1949. A coexistência de elementos chineses e portugueses reflete-se na legislação híbrida, na administração pública e no bilinguismo institucional, reforçando a singularidade de Macau no âmbito do modelo "um país, dois sistemas".

1.1 O bilinguismo em Macau

O bilinguismo em Macau apresenta, segundo Leong (2012), características singulares, moldadas por sua história e pelo enquadramento jurídico da RAEM. Conforme estipulado pela Lei

Básica (1993), no seu Artigo 9.º, o chinês e o português são definidos como "línguas formais", uma designação deliberada que evita a conotação de soberania estatal associada ao termo "língua oficial". Na prática, observa-se uma clara hierarquia linguística: o chinês, língua materna de 94% da população, predomina no cotidiano, enquanto o português, falado por apenas 3,1%, exerce sobretudo um papel técnico e suplementar, especialmente no domínio jurídico. Leong (2012) informa que as leis são originalmente redigidas em português e, posteriormente, traduzidas para o chinês, o que acarreta desafios significativos, como traduções literais e estruturas sintáticas híbridas (nem portuguesas nem chinesas) que dificultam a compreensão por parte dos sinófonos. Além disso, a escassez de profissionais bilíngues qualificados compromete a precisão e a equivalência jurídica nas traduções, contribuindo para uma assimetria funcional entre as duas línguas.

Uma "língua oficial" é aquela expressamente reconhecida por uma constituição ou legislação como meio obrigatório de comunicação no funcionamento das instituições públicas, abrangendo áreas como a administração, a justiça e a educação (Leong, 2012). Seu estatuto legal confere-lhe primazia em documentos oficiais, atos governamentais e decisões judiciais, além de servir como critério para a resolução de conflitos linguísticos em contextos multilíngues. A oficialização de uma língua implica não apenas seu uso institucional, mas também o compromisso do Estado em garantir seu ensino, promoção e acesso igualitário. Um exemplo emblemático é o caso de Singapura, onde o inglês, o chinês, o malaio e o tâmil são línguas oficiais por força constitucional, refletindo a diversidade étnico-linguística do país e assegurando representatividade equitativa nas esferas públicas.

O termo "língua formal", tal como empregado na *Lei Básica de Macau* (1993), designa, segundo Leong (2012), uma categoria linguística específica cujo uso é obrigatório nas esferas executiva, legislativa e judicial da administração local, embora sem implicar a conotação política de soberania estatal associada à noção de "língua oficial". Ainda que esse conceito não seja comum na linguística, ele se aproxima funcionalmente da ideia de "língua oficial" em determinados contextos institucionais, como nos documentos das Nações Unidas. Em Macau, a escolha terminológica reflete um compromisso histórico e diplomático com Portugal, ao mesmo tempo em que evita tensionamentos simbólicos com a soberania chinesa, conferindo ao português e ao chinês um estatuto administrativo equiparado, mas com implicações políticas cuidadosamente delimitadas.

O impacto do bilinguismo jurídico em Macau manifesta-se de forma aguda na linguagem jurídica e nos complexos procedimentos de tradução jurídica que envolvem harmonização terminológica, adaptação cultural e validação institucional entre chinês e português. Atualmente, a direção predominante da tradução é do português para o chinês, já que as leis continuam a ser originalmente redigidas em português. Esse modelo gera textos jurídicos em chinês que frequentemente mantêm estruturas gramaticais portuguesas, o que compromete sua clareza e dificulta a compreensão por parte da população sinófona. Entre os principais problemas identificados estão as traduções literais, nas quais termos jurídicos em chinês não refletem o uso comum da língua, gerando ambiguidade e afastamento do leitor. Além disso, a forte dependência do português obriga os profissionais do direito a dominarem essa língua para interpretar corretamente os textos jurídicos, mesmo em uma sociedade majoritariamente chinesa. Como soluções, propõe-se a inversão do fluxo de tradução, redigindo originalmente as leis em chinês, e o investimento na

formação de tradutores jurídicos bilíngues, a fim de assegurar maior precisão terminológica e acessibilidade jurídica.

Nesse contexto, o projeto "Tradução jurídica em contexto" (TraJeC) propõe a análise sistemática de textos normativos em chinês e português por meio da construção de um corpus paralelo anotado, fundamentado em traduções oficiais e tecnologias de processamento de linguagem natural. A partir de dados compilados do site da Imprensa oficial de Macau, o corpus é cuidadosamente alinhado ao nível de frase, com anotações linguísticas que possibilitam o mapeamento de estruturas lexicais, sintáticas e semânticas, conforme as abordagens de Cheng e Sun (2021) e Gao (2021). Combinando rigor metodológico e aplicabilidade prática, o TraJeC oferece uma contribuição relevante tanto para a resolução de desafios locais quanto para o avanço dos estudos comparados em contextos jurídicos multilíngues.

1.2 Linguística de corpus paralelo

Com base nos critérios metodológicos propostos por Sardinha (2000) para construção de corpora linguísticos, este trabalho compila corpora paralelos jurídicos chinês-português que viabilizam a análise comparativa de textos legais no sistema bilíngue de Macau, atendendo aos requisitos de representatividade e especificidade terminológica do domínio jurídico.

Nosso corpus paralelo de estudo, o "Corpus Jurídico de Macau", que alinha textos jurídicos em português e chinês, ilustra os princípios teóricos de Lefer (2020) sobre corpora direcionais. Cada lei, como a Lei n.° 27/2024 (em português) e sua correspondente 第 27/2024 號法律 (em chinês), é meticulosamente alinhada em segmentos (frase), garantindo a harmonização terminológica e equivalência jurídica entre os sistemas legais. Os metadados detalhados, como o número do Boletim Oficial (B.O. n.°: 1, I Série, 2025/01/06) e datas de publicação, reforçam a confiabilidade do corpus para análises contrastivas. Este recurso não apenas facilita a tradução especializada e a interpretação de nuances legais, mas também documenta o período histórico de Macau pós-1999, preservando a evolução legislativa em contextos bilíngues. O corpus evidencia que a bidirecionalidade relativa (português ← → chinês), condicionada por assimetrias linguísticas e práticas de tradução, aliada ao alinhamento automático rigoroso, é fundamental para aplicações práticas, como a harmonização de normas e a garantia de equidade jurídica em sociedades multiculturais.

1.3 Descrição do corpus

O corpus foi compilado a partir do site da Imprensa Oficial (IO) do Governo da RAEM, seguindo a hierarquia informacional definida no mapa de navegação do portal, que organiza o conteúdo em múltiplos níveis. Os textos legislativos selecionados estão localizados no nível 3 dessa estrutura, correspondente a uma camada específica dedicada à organização de documentos normativos. Essa estratificação permite um acesso lógico e sistematizado ao material jurídico, alinhando-se ao planejamento dos desenvolvedores do site para facilitar a busca e consulta pelos usuários.

Tabela 1: Os três níveis do site da Imprensa Oficial (IO) do Governo da RAEM

Mapa	Descrição	Acesso ao portal em português	Acesso ao portal em chinês
Nível 1	*Home*	https://www.io.gov.mo/pt/home/	https://www.io.gov.mo/cn/home/
Nível 2	Legislação	https://www.io.gov.mo/pt/legis	https://www.io.gov.mo/cn/legis
Nível 3	Leis	https://www.io.gov.mo/pt/legis/list/a/?d=46	https://www.io.gov.mo/cn/legis/list/a/?d=46

Fonte: Autor (2025)

O site da Imprensa Oficial (IO, 2025) apresenta na sua página inicial (*Home*, nível 1) em português e em chinês, um design institucional completo com cabeçalho contendo logotipo e opções de navegação, menu principal organizado em sete categorias como "Boletim Oficial", "Legislação", "Notícias", "Entidades", "Online", "Ligações" e "Instruções". A área central é dedicada a notícias e anúncios oficiais recentes, acesso facilitado ao Boletim Oficial da RAEM em ambas as séries, a uma ferramenta de pesquisa para localização de documentos, a links direcionados a outros serviços governamentais de Macau e uma opção para alternar entre dois idiomas (chinês tradicional, português), cumprindo assim sua função primordialmente institucional como portal para acesso às publicações oficiais, legislação do governo macaense e informações sobre os serviços prestados pela Imprensa Oficial.

A seção dedicada à "Legislação" (Nível 2) no site da Imprensa Oficial em português, oferece aos usuários um robusto sistema de pesquisa e consulta legislativa que inclui acesso a diversas categorias jurídicas ("Legislação", "Direito Internacional", "Legislação anterior à RAEM", "Lei Básica e Outros"), ferramentas de filtragem por tipo de documento (Leis, Regulamentos Administrativos, Ordens Executivas, Despachos, entre vários outros), interface de busca textual completa, além de links para recursos legislativos complementares, funcionando efetivamente como um abrangente repositório e ferramenta de consulta da legislação oficial macaense que permite aos cidadãos, profissionais do direito e demais interessados acessarem com facilidade todas as normas legais vigentes na RAEM.

Na página que apresenta "Leis" (Nível 3), encontra-se uma listagem ordenada de leis filtradas conforme o parâmetro cronológico padrão (a partir da mais recente) que representa uma categorização específica por tipo de documento, o ano de publicação, a área temática dentro do sistema legislativo macaense, apresentando uma interface organizada com tabelas contendo informações essenciais sobre cada lei, como número, data de promulgação, título oficial, status de vigência e links de acesso ao texto integral, além de ferramentas para refinamento de pesquisa, ordenação alternativa dos resultados e navegação entre diferentes categorias relacionadas, servindo assim como um importante recurso de consulta para a sociedade macaense.

A tabela seguinte apresenta uma amostra da organização do corpus bilíngue composto pelas 100 leis mais recentes, nesta data, com textos correspondentes em português e chinês.

Cada linha exibe para os dois corpora: (1) a "Descrição" (Título completo da lei em português, ex.: "Lei n.º 27/2024, Adaptação e integração de leis..."); (2) o "B.O. n.º:" (Referência ao Boletim Oficial), indicando a edição, a série do B.O. e (3) a data de publicação (ex.: "B.O. n.º: 1, I Série, 2025/01/06"). Uma estrutura similar se aplica ao corpus em chinês: (1) 摘要 (Resumo): (Título completo da lei em chinês (ex.: "第 27/2024 號法律, 一九九四年至 ……"); (2) 《公報》編號 (Número do Boletim Oficial) e (3) a data de publicação (ex.: "《公報》第 I 期, 第一組, 2025/01/06").

Tabela 2: Amostra da organização do corpus bilíngue

CORPUS - PORTUGUÊS		CORPUS - CHINÊS	
Descrição	**B.O. n.º:**	摘要	《公報》編號:
Lei n.º 27/2024, Adaptação e integração de leis e decretos-leis publicados entre 1994 e 1999.	B.O. n.º: 1, I Série, 2025/01/06	第 27/2024 號法律, 一九九四年至一九九九年公佈的若干法律及法令的適應化及整合。	《公報》第 I 期, 第一組, 2025/01/06
[...]	[...]	[...]	[...]

Fonte: Autor (2025)

A estrutura do corpus desenvolvida no projeto TraJeC foi definida para assegurar a melhor precisão e utilidade analítica, combinando paralelismo textual, rigor documental e identificação temática. Cada lei é apresentada, conforme consta no site, em pares perfeitamente alinhados (português e chinês) permitindo a correspondência direta entre os textos e facilitando análises comparativas detalhadas. Além disso, a estrutura do corpus integra metadados essenciais, como o número do Boletim Oficial, série e data de publicação, garantindo a rastreabilidade jurídica e a verificação das fontes originais.

2. Metodologia

2.1 Compilação de corpora

A compilação do corpus jurídico macauense privilegiou a representatividade funcional em detrimento da proporcionalidade quantitativa, considerando as especificidades do ordenamento jurídico híbrido de Macau. A seleção textual abrange domínios legislativos estratégicos, como o direito administrativo, a regulamentação financeira, a legislação sobre jogos de azar e as normativas de imigração, que refletem tanto a herança jurídica lusófona quanto as adaptações necessárias ao contexto chinês contemporâneo. Esta abordagem metodológica reconhece que a representatividade em corpora jurídicos bilíngues deve ser avaliada pela capacidade de capturar padrões terminológicos e estruturas discursivas características da tradução jurídica institucional, em vez de uma distribuição estatística uniforme entre categorias temáticas.

2.1.1 Coleta automatizada de textos legislativos

O site da Imprensa Oficial destaca-se pela clareza, organização e simplicidade técnica. Trata-se de uma aplicação web estática, que cumpre de maneira eficiente sua função de disponibilizar legislações ao público por meio de uma estrutura funcional e enxuta, facilitando também a coleta automatizada de dados com scripts em Python[1]. A construção do corpus paralelo de textos jurídicos

[1] Os scripts e os corpora criados nesta pesquisa encontram-se disponíveis no repositório de acesso aberto Figshare (Miroir, 2025).

de Macau envolveu o desenvolvimento de scripts Python personalizados para extração automatizada de legislações em português e chinês disponíveis no portal da Imprensa Oficial (2025).

O script de extração automatizada de leis em português realiza as seguintes operações:

1) Define o endereço principal de coleta e cria uma pasta local (Corpus_leis_pt) destinada a armazenar os textos extraídos.
2) A partir da página principal de listagem de leis, o script identifica e coleta todos os links que correspondem ao padrão típico de nomenclatura legislativa ("Lei n.º XX/YYYY"), utilizando expressões regulares (regex) para garantir a precisão.
3) Em caso de redirecionamentos automáticos, o script ajusta a URL para garantir o acesso direto ao conteúdo final, aumentando a confiabilidade da coleta.
4) Para cada URL, o script utiliza a biblioteca *BeautifulSoup* para extrair o corpo principal do texto jurídico. Estratégias alternativas de extração são aplicadas caso o formato da página varie, garantindo a robustez do processo.
5) Cada lei extraída é salva como um arquivo .txt (texto simples) com codificação UTF-8, em um formato de nomenclatura padronizado (Lei_XX-YYYY_pt.txt), assegurando organização e rastreabilidade dentro do corpus.
6) Para respeitar a integridade do servidor, o script inclui pausas (time.sleep) entre as requisições.

Este processo gera um corpus textual limpo e estruturado, adequado para análises linguísticas e estudos comparativos no campo da linguística de corpus jurídica. A automação não apenas otimiza o tempo de coleta, mas também assegura a sistematicidade na formação do banco de dados, o que é fundamental para pesquisas quantitativas e qualitativas em linguística.

O script de extração automatizada de leis em chinês realiza as seguintes operações e apresenta particularidades técnicas específicas:

1) Devido à diversidade de codificações possíveis (UTF-8, Big5, GB18030, GBK), o script utiliza a biblioteca *chardet* para identificar dinamicamente a codificação correta de cada página antes da extração do conteúdo. Isso assegura a preservação adequada dos caracteres chineses.
2) Os cabeçalhos HTTP (headers) foram personalizados para simular um navegador com preferência linguística chinesa (Accept-Language: zh-CN), o que melhora a resposta do servidor para conteúdos localizados.
3) O script adapta as expressões regulares para reconhecer os títulos de leis no formato específico utilizado em chinês ("第 XX/YYYY 號法律"), garantindo a seleção precisa dos textos.
4) Os textos extraídos são salvos em UTF-8 com BOM (utf-8-sig), o que amplia a compatibilidade em softwares de processamento de texto que manipulam chinês tradicional ou simplificado.

A necessidade de desenvolver dois scripts distintos para a coleta de leis em português e chinês fundamenta-se em diferenças técnicas e linguísticas essenciais. Enquanto as páginas em português apresentam codificação homogênea em UTF-8, as páginas em chinês exigem a detecção de múltiplas codificações, como Big5 e GB18030. Além disso, a estrutura textual dos documentos, os padrões de nomenclatura e as configurações de comunicação com o servidor variam entre os idiomas, o que demanda a elaboração de expressões regulares e cabeçalhos de requisição específicos para cada contexto. A preservação da integridade dos caracteres chineses também requer estratégias de arquivamento diferenciadas. A separação dos processos, portanto, garante a formação de corpora paralelos limpos, confiáveis e linguisticamente fiéis, assegurando a robustez metodológica necessária para o alinhamento automatizado frase por frase e lei por lei.

A extração automatizada resultou, conforme destacado no quadro seguinte, na criação de dois repositórios organizados, "Corpus_leis_pt" e "Corpus_leis_zh" [2] (Miroir, 2025), correspondentes, respectivamente, às legislações em português e em chinês. Em cada pasta, os arquivos de texto foram sistematicamente nomeados de acordo com um padrão padronizado (por exemplo, "Lei_01-2021_pt.txt e Lei_01-2021_zh.txt"), o que assegura a correspondência direta entre os documentos bilíngues. Essa organização facilita o alinhamento paralelo entre os textos, essencial para análises comparativas, estudos de tradução jurídica e investigações de correspondência terminológica. Além disso, a padronização dos nomes de arquivos permite a automação das próximas tarefas de processamento, como o alinhamento automático de segmentos e as análises linguísticas assistidas por ferramentas de corpus.

Tabela 3: Amostra da nomeação dos artigos dos corpora

Corpus_leis_pt	Corpus_leis_zh
Lei_01-2021_pt.txt	Lei_01-2021_zh.txt
Lei_01-2022_pt.txt	Lei_01-2022_zh.txt
Lei_01-2023_pt.txt	Lei_01-2023_zh.txt
[...]	[...]

Fonte: Autor (2025)

Para garantir a organização e a reprodutibilidade do projeto, foi criado um repositório *Python_Scripts*, destinado a hospedar os ambientes virtuais (venv) utilizados na execução dos scripts de extração e processamentos de dados. O ambiente virtual dedicado à coleta de textos foi denominado *scrap*, isolando todos os pacotes e respectivas versões necessários para essa tarefa específica, como *requests, BeautifulSoup4* e *chardet*. Essa estrutura assegura que as dependências sejam gerenciadas de forma controlada, sem interferir em configurações globais do sistema. Além de proporcionar maior estabilidade, essa organização facilita a manutenção e atualização dos scripts e permite a replicação fiel do ambiente de desenvolvimento para outros pesquisadores que venham a utilizar ou expandir o corpus.

[2] De acordo com a norma ISO 639, "zh" é o código para o idioma chinês. As etiquetas de idioma geralmente consistem em uma ou mais subtags separadas por hifens. A subtag primária normalmente representa o idioma, e subtags adicionais podem especificar a região, o script ou outras variantes, por exemplo: Chinês (China): zh-CN - Chinês simplificado usado no interior da China; Chinês (Hong Kong): zh-HK - Chinês tradicional usado em Hong Kong; Chinês (Taiwan): zh-TW - Chinês tradicional usado em Taiwan; **Chinês (Macau): zh-MO - Chinês tradicional usado em Macau**; Português (Brasil): pt-BR - Português como falado no Brasil ; Português (Portugal): pt-PT - Português como falado em Portugal. Disponível em: https://centus.com/blog/iso-language-codes.

2.1.2 Alinhamento automatizado de textos legislativos em português e em chinês

Durante a tentativa de alinhamento frase por frase entre os textos jurídicos em português e em chinês, foram encontradas dificuldades significativas ao utilizar ferramentas de alinhamento de uso geral *online*, gratuitas, como *Matecat Aligner, Wordfast Anywhere Aligner* (nas versões gratuita e comercial) e o *Web Align Toolkit*, com testes realizados nos quatro modelos disponíveis (*LF Aligner, YASA, JAM* e *Alinéa Lite*). Em todos os casos, os resultados mostraram-se insatisfatórios, seja pela associação incorreta de segmentos, seja pela geração de alinhamentos vazios. Essas limitações estão associadas ao fato de que tais ferramentas, geralmente otimizadas para línguas românicas e outros sistemas de escrita alfabéticos, apresentam baixo desempenho ao lidar com modos de escrita não alfabéticos, como o chinês, que exige abordagens específicas para segmentação e correspondência textual.

Um desempenho relativamente mais satisfatório foi obtido com o *YouAlign* (em sua versão de testes gratuitos limitados), que apresentou maior precisão no alinhamento de segmentos curtos e estruturalmente distintos. Diante dessas limitações, a solução definitiva foi a adoção do Bertalign, que viabilizou a construção de um corpus paralelo mais confiável. O método adotado não apenas realiza alinhamentos frase por frase com alta precisão, mas também permitiu processar todos os arquivos do corpus inteiro de forma automatizada via scripts em Python, uma funcionalidade indisponível nas ferramentas de uso geral testadas.

O algoritmo Bertalign foi desenvolvido por Lei Liu e Min Zhu (2022), pesquisadores da *School of Foreign Languages* da *Yanshan University*, China. Esse algoritmo utiliza técnicas avançadas de inteligência artificial (utilizando um modelo de *embeddings* - LaBSE) sendo, portanto, capaz de identificar correspondências semânticas, ou seja, frases que expressam o mesmo conteúdo, mesmo que redigidas em línguas estruturalmente muito diferentes, como o português e o chinês. O alinhamento ocorre frase por frase, de forma totalmente automatizada, o que garante um nível de precisão e adequação muito superior aos métodos tradicionais de alinhamento, que geralmente se baseiam apenas no tamanho dos segmentos ou na posição de sinais de pontuação.

Diferentemente das ferramentas de alinhamento de uso geral disponíveis online, o Bertalign exige a instalação em máquina local a partir de seu repositório no GitHub e a configuração de ambientes Python com bibliotecas específicas, como PyTorch e HuggingFace Transformers, entre outras. Embora sua execução requeira conhecimentos técnicos, a ferramenta oferece resultados de alinhamento de alta qualidade e viabiliza a organização automática de corpora inteiros.

Na máquina local, foi criada uma pasta nomeada "Github" para armazenar o código-fonte do Bertalign, clonado de seu repositório público no GitHub (bfsujason, 2022), e as dependências necessárias ao projeto. Para sua execução, foi criado um ambiente virtual dedicado (ba) na pasta "Python_Scripts", isolando as dependências específicas do alinhador (como *sentence-transformers*). Essa estrutura assegura o gerenciamento controlado de pacotes, evitando conflitos com configurações globais da máquina e outros algoritmos do projeto. Além de facilitar atualizações e manutenção, esse ambiente permite a replicação precisa por outros pesquisadores, garantindo reprodutibilidade técnica e profissional na expansão do corpus.

Para construir um corpus paralelo confiável entre o português e o chinês, foi desenvolvido um script em Python "run_bertalign.py" que utiliza o Bertalign para o alinhamento automático de frases e realiza as seguintes operações:

1) Antes de iniciar o alinhamento, o script se conecta respectivamente às pastas dos corpora compilados na etapa anterior: "Corpus_leis_pt" e "Corpus_leis_zh" (Miroir, 2025). A padronização dos nomes dos arquivos (por exemplo, Lei_01-2021_zh.txt e Lei_01-2021_pt.txt) permite que o script identifique automaticamente os pares de documentos a serem alinhados.
2) Para cada par de arquivos do corpus chinês (Corpus_leis_zh) e português (Corpus_leis_pt), o script abre o arquivo e lê seu conteúdo. Esse passo garante que o alinhador receba apenas os textos corretos, pronto para processamento.
3) O script inicializa o alinhador Bertalign para cada par de textos, permitindo a comparação automática entre frases.
4) Após o alinhamento das frases de cada par de arquivos, o script gera uma planilha *Excel* estruturada com duas colunas: a primeira contendo as frases em chinês e a segunda, as respectivas frases correspondentes em português, seguindo a nomenclatura padronizada "aligned_sentences_Lei_XX-YYYY_zh-pt.xlsx", o que facilita a identificação e a organização dos dados. Por exemplo, o arquivo "aligned_sentences_Lei_01-2021_zh-pt.xlsx" reúne todas as frases da Lei 01/2021, alinhadas lado a lado.
5) O script cria, na primeira rodada, uma pasta "Corpus_aligned_sentences_zh-pt" (Miroir, 2025) que recebe todas as planilhas *Excel* das 100 leis alinhadas.

Tabela 4: Amostra da nomeação dos arquivos alinhados

Corpus_aligned_sentences_zh-pt
aligned_sentences_Lei_01-2021_zh-pt.xlsx
aligned_sentences_Lei_01-2022_zh-pt.xlsx
aligned_sentences_Lei_01-2023_zh-pt.xlsx
[...]

Fonte: Autor (2025)

Os arquivos com os textos alinhados são compostos por duas colunas: a da esquerda apresenta o texto em chinês (zh) e a da direita traz o texto correspondente em português (pt), segmento por segmento.

Os alinhamentos foram executados em uma máquina DELL com processador Intel Core i7 de 11ª geração, com frequência base de 2.80 GHz e 16 GB de memória RAM. Esta configuração demonstrou-se apropriada para a execução do Bertalign nas tarefas de alinhamento de corpora de dimensão moderada. A velocidade de processamento alcançou uma média de 15 textos legislativos por hora, resultando em um tempo total de 6 horas e 30 minutos para realizar o alinhamento completo dos 100 documentos que compõem os corpora português e chinês.

Tabela 5: Amostra de conteúdo de um arquivo alinhado

aligned_sentences_Lei_01-2021_zh-pt.xlsx	
zh	**pt**
澳門特別行政區	REGIÃO ADMINISTRATIVA ESPECIAL DE MACAU
公報 - 第一組	BOLETIM OFICIAL - I SÉRIE
法規：	Diploma:
第 1/2021 號法律	Lei n.º 1/2021
公報編號：	BO N.º:
5/2021	5/2021
刊登日期：	Publicado em:
2021.2.1	2021.2.1
版數：	Página:
101-105	101-105
從事科技創新業務企業的稅務優惠制度。	Regime de benefícios fiscais para as empresas que exerçam actividades de inovação científica e tecnológica.
葡文版本	Versão Chinesa
相關法規	Diplomas relacionados
:	:
第 24/2024 號法律	Lei n.º 24/2024
- 核准《稅務法典》。	- Aprovação do Código Fiscal.
第 21/78/M 號法律	Lei n.º 21/78/M
- 核准超額純利稅章程—— 取消一九六四年六月二日第 1635 號、一九六五年二月十三日第 1659 號、一九六五年六月十二日第 1668 號、一九六六年九月十日第 1718 號、一九六九年三月一日第 1787 號、一九七零年三月十四日第 1814 號等立法條例及三月十二日第 7/77/M 號法令	- Aprova o Regulamento do Imposto Complementar de Rendimentos. — Revoga os Diplomas Legislativos n.os 1 635, de 2 de Junho de 1964, 1 659, de 13 de Fevereiro de 1965, 1 668, de 12 de Junho de 1965, 1 718, de 10 de Setembro de 1966, 1 787, de 1 de Março de 1969, 1 814, de 14 de Março de 1970, e o Decreto-Lei n.º 7/77/M, de 12 de Março.
[...]	[...]

Fonte: Autor (2025)

2.2 Anotação dos corpora

A anotação de corpus, segundo John Newman e Christopher Cox (2020), refere-se ao processo de adicionar informações linguísticas ou metalinguísticas a um corpus, enriquecendo-o com dados estruturados que facilitam análises específicas. Isso inclui categorizar palavras por classes gramaticais (POS tagging), reduzir termos a suas formas base (lematização), analisar estruturas sintáticas, atribuir rótulos semânticos e outras camadas de informação. A anotação é geralmente automatizada para lidar com grandes volumes de texto, mas também pode envolver ajustes manuais para garantir precisão. Seu objetivo principal é tornar padrões linguísticos mais acessíveis e viabilizar pesquisas eficientes, especialmente em análises de corpora paralelos alinhados, por meio de ferramentas como o Antconc e o AntPconc (específico para corpora paralelos).

2.2.1 Anotação de palavras por classes gramaticais e lematização: Corpus_leis_pt

Para o processamento linguístico do corpus de leis em português, foi utilizado o spaCy, uma biblioteca de código aberto para Processamento de língua natural (PLN). O *pipeline* em português foi configurado para otimização em CPU, garantindo a eficiência computacional mesmo em ambientes sem aceleradores gráficos (GPU). O *morphologizer* (etiquetador) é responsável por anotar informações morfossintáticas com os *Universal POS Tags* (UPOS), padronizando as principais categorias gramaticais (ex: VERB, NOUN) para as línguas do mundo, o que facilita comparações interlinguísticas. O *lemmatizer* permite a redução das palavras flexionadas às suas formas canônicas (lemas), como "foi" → "ser". Essa função é essencial para a normalização e a análises semânticas das línguas flexionadas com o português. O modelo spaCy empregado é o "pt_core_news_lg" (versão "large", ou seja, com cerca de 500 mil vetores de palavras), pré-treinado em um corpus diversificado de textos em português (notícias e mídia).

Essas especificidades do spaCy são implementáveis via programação em Python. Contudo, para usuários que não desejam trabalhar diretamente com código, o TagAnt (Anthony, 2024c) destaca-se como uma ferramenta acessível. Desenvolvido por Laurence Anthony, esse software gratuito foi projetado para segmentação de texto e etiquetagem gramatical (*POS tagging*) em múltiplos idiomas, executando os modelos do spaCy em seu *backend*. O programa automatiza essas tarefas avançadas de processamento linguístico, como as análises morfossintáticas e a lematização, de forma transparente para o usuário, dispensando assim conhecimentos técnicos em programação. Sua interface amigável permite que pesquisadores e linguistas de corpus processem grandes volumes de texto com eficiência, garantindo alta precisão na anotação graças à integração com os recursos do spaCy. Essa combinação de facilidade de uso e robustez técnica torna o TagAnt uma ferramenta valiosa para análises de corpora multilíngues e estudos linguísticos que demandem confiabilidade e escalabilidade.

O *Language Model Manager* do TagAnt possibilita a configuração personalizada do modelo linguístico (spaCy - "pt_core_news_lg") e dos três níveis de processamento selecionados (word+pos+lemma), para executar simultaneamente três etapas essenciais: a tokenização, a análise morfossintática (*morphologizer UPOS*) e a lematização. Para este estudo, integrou-se à ferramenta TagAnt a pasta "Corpus_leis_pt", que contém os 100 arquivos de textos jurídicos em português para processamento e carregou automaticamente todos os documentos. Graças à otimização para CPU e à eficiência computacional do spaCy, o processamento completo foi executado com rapidez, demonstrando a viabilidade da ferramenta para análise de grandes volumes de dados linguísticos mesmo em *hardware* convencional. O TagAnt cria uma pasta "tagged" para receber o corpus devidamente etiquetado.

2.2.2 Anotação de palavras por classes gramaticais: Corpus_leis_zh

A anotação de palavras por classes gramaticais e a lematização são, de fato, tarefas essenciais no processamento de toda linguagem natural. No entanto, aplicar esses procedimentos à língua chinesa apresenta desafios devido às características dessa língua. A ambiguidade de classe gramatical é um dos principais obstáculos, pois muitas palavras chinesas podem pertencer a diferentes classes

gramaticais dependendo do contexto. Por exemplo, uma palavra pode ser um verbo em uma frase e um substantivo em outra, complicando a categorização automática. Além disso, o chinês é uma língua isolante, onde as palavras não mudam de forma para indicar tempo, número ou caso, o que dificulta a identificação da função gramatical de uma palavra apenas pela sua forma. A interpretação correta de uma palavra frequentemente depende do contexto, exigindo que ferramentas de PLN analisem frases ou parágrafos inteiros para categorizar palavras corretamente.

Quanto à lematização, ela tem, por conseguinte, uma utilidade limitada no chinês devido às características da língua. Diferentemente de idiomas como português ou inglês, o chinês não possui flexão morfológica, os verbos não se conjugam e os substantivos não mudam para indicar o plural ou o gênero. Assim, a maioria das palavras já está, no uso, em sua forma canônica, tornando a lematização redundante e desnecessária. Para o processamento automático do chinês, tarefas como segmentação de palavras (tokenização), identificação de compostos léxicos e desambiguação semântica são muito mais necessárias do que a busca por lemas.

O *Language Model Manager* do TagAnt (Miroir, 2024a) possibilita a configuração personalizada do modelo linguístico (spaCy - "zh_core_news_lg") e dos dois níveis de processamento selecionados (word+pos), para executar simultaneamente duas etapas essenciais: a tokenização e a análise morfossintática (*morphologizer UPOS*). Podemos observar que a qualidade da tokenização é fundamental para separar as palavras nas frases chinesas. Segundo o site spaCy, acurácia da tokenização é de 96% para o chinês e de 100% para o português (tarefa muito mais fácil nesta língua).

Para esse estudo, integrou-se à ferramenta TagAnt a pasta "Corpus_leis_zh", que contém os 100 arquivos de textos jurídicos em chinês para processamento e carregou automaticamente todos os documentos. O TagAnt processa o "Corpus_leis_zh" e cria uma nova pasta denominada "tagged" para armazenar os arquivos etiquetados.

Devido à possibilidade de executar o processamento (word+pos_tag), isto é, a tokenização e a análise morfossintática com etiquetas (tags) mais detalhadas e específicas à língua chinesa (TAG-zh)[3], por exemplo: "*BA 把 in ba-construction*" (indica que o objeto está sendo manipulado ou afetado pelo verbo), *"DEC 的 in a relative clause*" (usado para descrever algo, como adjetivos ou posse). O TagAnt processa o "Corpus_leis_zh" e cria uma nova pasta denominada "tagged" para armazenar os arquivos resultantes do processamento.

2.3 Análise monolíngue dos corpora com *Antconc*

O Antconc (Anthony, 2024a) é um software multiplataforma usado em linguística de corpus, desenvolvido por Laurence Anthony, professor na Faculdade de Ciência e Engenharia da Waseda University, no Japão. Ele é amplamente utilizado por pesquisadores em (socio)linguística, em tradução e em ciências sociais para tarefas como a contagem de frequência de palavras, a análise de co-ocorrência (*collocates*), de agrupamentos (*clusters, n-grams*), de concordância, entre outras. O software é conhecido por sua facilidade de uso e não requer conhecimentos avançados de estatística nem de programação.

[3] Conjunto de etiquetas Penn Treebank para chinês.

2.3.1 Preparação do corpus português

O Antconc versão 4 introduziu uma importante melhoria ao incorporar o sistema de banco de dados SQLite para lidar com corpora maiores e mais complexos, sem comprometer seu desempenho. O corpus português etiquetado com UPOS[4], denominado "Corpus_leis_pt_pos" (Miroir, 2025), no formato texto simples (arquivos .txt) e com codificação de caracteres UTF-8, preparado nas etapas anteriores, é incorporado na ferramenta Antconc por meio do "*Corpus Manager*" (Miroir, 2024b). Este processo transforma automaticamente o corpus em um banco de dados *SQLite*, realizando a devida indexação das etiquetas UPOS, o que possibilita a execução de pesquisas linguísticas avançadas posteriormente. Após carregar o corpus no Antconc e verificar a importação (observando os resultados da ferramenta "Word", anteriormente denominada "Wordlist"), as configurações de exploração das etiquetas (tags) devem ser verificadas tanto globalmente (*Global Settings* > *Tags*) quanto localmente (*Tool Settings* > "Ferramenta escolhida" > *Tags*). Algumas buscas simples combinando palavras e etiquetas UPOS podem ser executadas para confirmar o funcionamento adequado do sistema de indexação e consulta no corpus (Miroir, 2024b). Concluídas estas verificações, o corpus formatado como banco de dados (.db) pode ser salvo na máquina com o nome "Corpus_leis_pt_pos.db" (Miroir, 2025), para futuras explorações e análises.

2.3.2 Preparação do corpus chinês

Da mesma forma de que para o corpus português, o corpus chinês etiquetado com as tags especificas para chinês (TAG-zh), denominado "Corpus_leis_zh_pos-tag" (Miroir, 2025), no formato texto simples (arquivos .txt) e com a codificação de caracteres UTF-8, preparado nas etapas anteriores, é incorporado na ferramenta Antconc por meio do *Corpus Manager*. A escolha do conjunto de tags específicas para chinês (TAG-zh) é justificada pela melhor acurácia da tokenização em comparação com a tokenização feita na etiquetagem com UPOS (0, p.17). Além disso, o corpus etiquetado com UPOS "Corpus_leis_zh_pos" está disponível para explorações similares com *Antconc*. Concluídas as verificações semelhantes à preparação do corpus português (ver **0**, acima), os corpora chineses (TAG-zh e UPOS), formatados como banco de dados (.db), podem ser salvos na máquina local com os nomes "Corpus_leis_zh_pos-tag.db" (TAG-zh) e "Corpus_leis_zh_pos.db" (UPOS), para serem explorados em futuras análises.

2.4 Análise bilíngue dos corpora

2.4.1 Execução local na máquina com *AntPconc*

AntPconc (Anthony, 2024b) é um software multiplataforma especializado na análise de corpus paralelos, desenvolvido por Laurence Anthony para auxiliar pesquisadores e tradutores no

[4] O conjunto de etiquetas Penn Treebank para português existe; no entanto, o *TagAnt* processa, sem justificativa aparente, o corpus com as mesmas etiquetas UPOS. Uma solução seria usar o modelo *spaCy* "Portuguese (pt_core_news_lg-3.8.0)" com um script em Python.

estudo de textos alinhados em diferentes línguas. Diferente do Antconc, que é voltado para análise monolíngue, o AntPconc permite extrair concordâncias bilíngues, identificar padrões de tradução e comparar estruturas linguísticas entre idiomas. Com suporte a buscas avançadas, visualização intuitiva e exportação de dados, é uma ferramenta essencial para linguistas de corpus, tradutores e terminólogos que trabalham com textos paralelos.

O AntPconc precisa ser alimentado por um corpus previamente alinhado, como o "Corpus_aligned_sentences_zh-pt" (arquivos .xlsx) preparado nas etapas anteriores (ver **0**, p. 9). Contudo, o corpus alinhado é composto por um corpus em chinês, "Corpus_leis_zh" (arquivos .txt), que deve ser preparado por meio de segmentação de palavras (tokenização) para que o AntPconc possa identificar as expressões de busca em chinês, conforme exposto anteriormente (0, p.12). A ferramenta TagAnt pode executar essa tarefa, utilizando a opção "Word" (token). O novo corpus gerado é salvo na máquina e nomeado como "Corpus_leis_zh-tkn" (arquivos .txt) (tkn para token) para, em seguida, ser alinhado com o corpus "Corpus_leis_pt" (arquivos .txt) utilizando a ferramenta Bertalign.

O novo corpus alinhado "Corpus_aligned_sentences_zh-tkn-pt" (arquivos .xlsx) não pode ser explorado diretamente pelo *AntPconc*, pois é composto por 100 planilhas *Excel*, uma para cada um dos textos (**0**, p. 9). As planilhas precisam ser agrupadas, por meio de um script em Python, em uma única planilha, para, em seguida, extrair e converter em texto simples (extensão .txt) a coluna em chinês, "Corpus_Leis_zh-tkn.txt" (4,3 MB, 70 mil linhas) e, da mesma forma, extrair e converter a coluna em português "Corpus_Leis_pt.txt" (4,9 MB, 70 mil linhas).

Ao abrir o *AntPconc*, a funcionalidade "Build/Edit corpus" abre o "Corpus Builder", onde é possível inserir os arquivos "Corpus_Leis_pt.txt" e "Corpus_Leis_zh-tkn.txt". Nesta etapa, a nomeação desses corpora é possível na ferramenta como "Corpus_Leis_pt" e "Corpus_Leis_zh-tkn", respectivamente, permitindo assim escolher a direcionalidade das pesquisas, seja em português ou em chinês (Miroir, 2024a).

Algumas buscas simples podem ser executadas, tanto em português quanto em chinês, para confirmar o funcionamento adequado do sistema de indexação, o alinhamento correto das frases e a consulta no corpus (Miroir, 2024b). Concluídas estas verificações, o corpus formatado como banco de dados (.db) pode ser salvo através da opção "*Save Corpus to Database*" na máquina local com o nome "Corpus_Leis_zh-tkn-pt.db" (52MB), para futuras explorações e análises.

A versão atual do AntPconc (versão 1.2.1) não permite explorar diretamente as etiquetas de categorização das classes gramaticais, como na ferramenta *Antconc*. Todavia, os procedimentos de preparação dos corpora alinhados e etiquetados podem ser executados sem problemas. As pesquisas podem ser feitas com os nomes das etiquetas (UPOS ou TAG-zh), consideradas como "tokens" ou "palavras". Esse tipo de busca pode evidenciar resultados instigantes.

2.4.2 Execução online no site da Sketch Engine

A ferramenta Sketch Engine é uma plataforma comercial online para análise linguística que oferece múltiplas funcionalidades complementares às do Antconc e do AntPconc, proporcionando aos pesquisadores um ecossistema linguístico completo com um período de teste gratuito de 30 dias. Diferentemente das ferramentas gratuitas de Anthony (2024a, 2024b, 2024c), o Sketch Engine

disponibiliza funcionalidades exclusivas como Word Sketches (perfis automáticos do comportamento gramatical e colocacional das palavras), Thesaurus (que identifica termos semanticamente relacionados com base em comportamentos distribucionais), WebBootCaT (para compilação automática de corpora a partir da web) e vários sistemas avançados para análise de corpora paralelos multilíngues.

A plataforma disponibiliza acesso imediato a centenas de corpora pré-compilados em mais de 90 idiomas, oferecendo processamento linguístico integrado (tokenização, lematização e etiquetagem automática), ferramentas de extração terminológica (OneClick Terms) e a capacidade de lidar com volumes massivos de dados em uma infraestrutura em nuvem, com a escalabilidade necessária para processamento muito rápido.

Tabela 6: Características quantitativas e qualitativas do corpus paralelo no Sketch Engine

Categoria	Corpus Chinês	Corpus Português	Análise Comparativa
ESTATÍSTICAS BÁSICAS			
Total de Tokens	820.362	889.188	Português 8,4% maior em volume bruto
Palavras (words)	668.149	700.753	Proporção palavras/tokens similar (~81,5% vs 78,8%)
Sentenças	70.851	79.792	Chinês tem sentenças mais longas (9,4 vs 8,8 palavras/sentença)
Documentos	1	1	Ambos compilados como documentos únicos
DIVERSIDADE LEXICAL			
Formas únicas	25.884	27.476	Português com 6,1% mais variação lexical
Tags gramaticais	35	174	Sistema português 5 vezes mais categorias (flexão verbal/nominal)
Lempos (lema + pos)	32.225	18.887	Chinês mostra maior polissemia (+70% combinações lema+tag)
RECURSOS MORFOLÓGICOS			
Gênero	N/A	18.464	Português exige marcação de gênero (masculino/feminino)
Análise morfêmica	N/A	18.886	Português com análise morfológica detalhada
Número	N/A	4	Flexão numérica obrigatória em português
Partículas	的, 地, 得	N/A	Chinês depende muito de partículas funcionais
ESTRUTURA			
Sentenças alinhadas	69.821	69.717	99,85% de correspondência (par quase perfeito)
Marcadores <s>	70.851	79.792	Estruturação sentencial completa em ambos
Marcadores <g>	N/A	129.821	Português com divisão adicional em parágrafos ou grupos
METADADOS			
Língua alinhada	Português	Chinês	Paralelo jurídico português-chinês

Fonte: Formatação dos dados fornecidos pelo Sketch Engine sobre cada corpus. Autor (2025)

O Sketch Engine representa um salto qualitativo fundamental em comparação com as soluções gratuitas disponíveis, especialmente para linguistas, tradutores, lexicógrafos e pesquisadores que necessitam de análises linguísticas baseadas em dependências sintáticas, visualizações avançadas, recursos multilíngues integrados e colaboração entre equipes. Seu modelo comercial é justificado tanto pela profundidade e sofisticação das análises linguísticas quanto pelos recursos computacionais implementados.

O corpus paralelo, composto de 100 planilhas *Excel*, fusionado em apenas uma planilha *Excel*[5], "Corpus_aligned_sentences_zh-tkn-pt.xlsx" (3.8 MB, 70 mil linhas), preparado para o AntPconc (0), foi enviado para o Sketch Engine para processamento do corpus paralelo para futuras pesquisas com as próprias ferramentas computacionais da plataforma (tempo de processamento de 12 minutos).

Doravante, o corpus paralelo está disponível no Sketch Engine ao público para livre consulta e exploração: "Corpus_Leis_Macau_zh" (Chinês tradicional) e "Corpus_Leis_Macau_pt" (Português). As características quantitativas e qualitativas do corpus paralelo, analisado por meio do Sketch Engine, são sintetizadas na tabela a seguir, acompanhadas de uma análise comparativa:

3. Resultados e discussão

3.1 Comparação de resultados de tokenização

Uma comparação da tokenização UPOS (word+pos) e a tokenização (word+pos_tag) permite visualizar as diferenças de resultados obtidos e, por conseguinte, de atribuição de etiquetas de categoria gramatical. A tabela seguinte compara essas duas abordagens de tokenização (segmentação de palavras) em textos chineses, demonstrando como diferenças na divisão de agrupamento de caracteres podem influenciar o significado e gerar impactos críticos em contextos jurídicos.

Tabela 7: Diferenças de sentido entre as tokenizações em chinês com UPOS e TAG-zh

Token. (UPOS)	Token. (TAG-zh)	Diferenças de sentido entre as tokenizações
上 項	上項	"上 項" pode ser interpretado como "item acima" ou "item superior", enquanto "上項" pode ser um termo consolidado (ex.: "item principal").
人 員	人員	"人 員" (pessoa + membro) pode ser mais genérico, enquanto "人員" é um termo consolidado para "pessoal" ou "funcionários".
兩 倍	兩倍	"兩 倍" (dois + vezes) e "兩倍" (dobro) têm o mesmo significado, mas a versão unificada é mais comum.
則該 項	則該項	"則該 項" (então + aquele + item) pode ser lido como separado, enquanto "則該項" é um termo jurídico consolidado ("o referido item").
創 新	創新	"創 新" (criar + novo) pode ser interpretado como um verbo + adjetivo, enquanto "創新" é o substantivo "inovação".

Fonte: Formatação dos dados elaborados com o apoio do modelo de linguagem chinês Deepseek. Autor (2025)

A primeira coluna "Token (UPOS)" apresenta uma tokenização baseada em critérios morfossintáticos, que tem tendencias a segmentar termos conforme unidades gramaticais individuais (ex.: "人 員" para "pessoa + membro"). Já a segunda coluna "Token (TAG-zh)" reflete a tokenização padrão do chinês, que trata expressões compostas como unidades únicas e consolidadas (ex.: "人員" para "pessoal"). A terceira coluna "Diferenças de sentido" detalha como essas variações na

[5] Na versão gratuita de teste do Sketch Engine, é permitido o uso de apenas um arquivo de textos alinhados (XLSX, TMX, XLIFF, TSV) para a criação de um corpus paralelo.

segmentação podem alterar a interpretação semântica, por exemplo, "創 新" (verbo + adjetivo) versus "創新" (substantivo: "inovação").

Essa comparação é particularmente relevante para linguistas de corpus, tradutores e profissionais do direito, pois evidencia como escolhas técnicas de processamento de texto podem ter repercussões tangíveis na exploração dos corpora com as ferramentas dedicadas e a interpretação das normas.

Esta tabela foi elaborada a partir da comparação entre dois sistemas de tokenização aplicados ao texto da Lei 1/2021 (arquivos Lei_1-2021_zh_pos.txt e Lei_1-2021_zh_pos_tag.txt). A análise semântica e interpretação dos resultados contou com o apoio do modelo de linguagem Deepseek (2023), um sistema avançado de processamento de língua chinesa disponível gratuitamente.

Vale ressaltar que para os falantes nativos de chinês, a diferença entre formas segmentadas (人 員) (leitura morfêmica) e termo lexicalizado (人員) geralmente passa despercebida na comunicação cotidiana, pois o significado global permanece inteligível em contexto. No entanto, essa distinção assume importância crítica em aplicações de Processamento de linguagem natural (PLN) e Tradução automática neural (TAN), particularmente em domínios especializados como textos jurídicos. Assim, se para humanos a variação equivale a grafias alternativas (como "guardachuva / guarda-chuva" em português), para máquinas trata-se de uma distinção operacional imprescindível, comparável à necessidade de ortografia padronizada em corretores automáticos. Esta dualidade revela a complexidade do chinês como língua isolante, onde limites entre palavra e morfema frequentemente desafiam tanto o processamento artificial quanto a análise linguística tradicional.

Na tabela a seguir, apresentam-se os resultados de acurácia referentes à categorização gramatical e à lematização no corpus *Corpus_leis_zh*, com base na avaliação dos modelos spaCy para o chinês (*zh_core_web_lg*) e para o português (*pt_core_news_lg*). A tokenização do texto em chinês demonstrou uma acurácia elevada (0,96) e um F-score de 0,93, indicando desempenho robusto na separação de palavras e sinais de pontuação, ainda que inferior ao desempenho observado no português (1,00 para ambas as métricas). A atribuição de etiquetas gramaticais detalhadas (TAG_ACC) obteve uma acurácia de 0,90 em ambos os idiomas, evidenciando consistência na identificação de classes específicas, como tempos verbais ou tipos de substantivos.

Tabela 8: Resultados de acurácia referentes à categorização gramatical e à lematização no corpus chinês e português

Função	Descrição	Acurácia zh	Acurácia pt	Comentário
TOKEN_ACC	Acurácia da tokenização	0.96	1.00	Avalia a correção da separação do texto em unidades básicas (tokens), como palavras e sinais de pontuação.
TOKEN_F	F-score da tokenização	0.93	1.00	Mede o desempenho geral na divisão do texto.
TAG_ACC	Acurácia da etiquetagem de classe gramatical (tags detalhadas)	0.90	0.90	Avalia a precisão das etiquetas gramaticais específicas atribuídas a cada palavra (por exemplo, diferenciar verbo no presente de verbo no passado).
SENTS_F	F-score da segmentação de sentenças (F-score)	0.75	0.94	Mede o desempenho geral na segmentação de sentenças.

POS_ACC	Acurácia da etiquetagem de classe gramatical (tags gerais - UPOS)	—	0.97	Avalia a precisão da atribuição de categorias gramaticais gerais (como substantivo, verbo, adjetivo).
LEMMA_ACC	Acurácia da lematização	—	0.97	Avalia a capacidade do modelo de identificar corretamente a forma base das palavras (ex.: "correram" → "correr").

Fonte: Formatação dos dados disponíveis no site da spaCy para os modelos chinês e português. Autor (2025)

Em contraste, a segmentação de sentenças apresentou um desempenho consideravelmente inferior no corpus chinês (F-score de 0,75), em comparação ao corpus em português (0,94), o que reflete a maior complexidade da identificação de fronteiras sentenciais no chinês escrito. Além disso, não foi reportada a acurácia para categorias gramaticais gerais (POS_ACC) nem para a lematização (LEMMA_ACC) no modelo chinês, ao passo que o modelo em português alcançou índices elevados de 0,97 para ambas as funções. Esses resultados evidenciam as diferenças estruturais e desafios técnicos inerentes ao processamento automático de línguas tipologicamente distintas.

A tokenização da língua chinesa representa até hoje um desafio significativo para o Processamento de linguagem natural (PLN) e para os sistemas de Tradução automática neural (TAN). No presente estudo, o tradutor DeepL, embora amplamente reconhecido como um dos mais avançados do mercado, não foi usado devido a problemas de tradução para o português resultantes de variações na tokenização do chinês. O tradutor automático da empresa chinesa Baidu, adotado nesta pesquisa, demonstrou maior robustez e menor sensibilidade às variações de tokenização na tradução do chinês para o português.

3.2 Exploração do corpus chinês e português com OneClick Terms

A ferramenta OneClick Terms (parcialmente disponível na versão de teste), integrada ao Sketch Engine, é especializada em extração automática de termos monolíngues e bilíngues, desenvolvida para apoiar tradutores e terminólogos na elaboração de glossários técnicos precisos. Com recursos como análise estatística avançada (Co-freq, logDice e Keyness), comparação de corpora paralelos e filtros inteligentes, a ferramenta identifica e valida pares de tradução relevantes, eliminando ruídos e garantindo consistência terminológica. O OneClick Terms agiliza a criação de glossários bilingues, aumentando a produtividade e a qualidade em áreas como jurídica, técnica e médica, conforme a amostra gerada a partir de nosso corpus paralelo instalado:

Tabela 9: Amostra de extração automática de termos bilíngues da ferramenta OneClick Terms

Source term	Target term	TAN - Baidu (comparativo)	Co-freq	L1 freq	L2 freq	Log Dice	Keyness
相關 物品	coisa conexa	itens relacionados	27	168	158	12,01	205,79
博彩 中介	promotor de jogo	intermediário de jogo	21	204	177	11,38	249,67
特别行政区	região administrativa	distrito político	31	203	203	11,14	248,45
金融 管理局	AMCM	gabinete de gestão financeira	46	374	335	11,05	456,90
公共 行政	administração pública	administração pública	15	115	79	10,99	141,18
工作 人	trabalhadores da administração	funcionários	6	126	20	10,98	154,59

證明 文件	documento de identificação	documentação	20	137	74	10,98	168,00
財政 年度	ano económico	exercício financeiro	17	227	178	10,73	277,71
行政 違法	infração administrativa	violação administrativa	45	437	468	10,67	533,69

Fonte: Formatação de termos bilíngues da ferramenta OneClick Terms do Sketch Engine. Autor (2025)

Para interpretar as métricas de extração automática de termos, vamos usar o exemplo da primeira linha da tabela, temos o par de termos "相關 物品" e "coisa conexa":

1) Co-freq (27): Indica que os termos aparecem juntos 27 vezes no corpus, mostrando uma relação direta entre eles.
2) L1 freq (168): O termo em chinês "相關 物品" aparece 168 vezes no total no texto-fonte (L1 - "Corpus_Leis_Macau_zh").
3) L2 freq (158): A tradução "coisa conexa" ocorre 158 vezes no texto-alvo (L2 - "Corpus_Leis_Macau_pt").
4) Log Dice (12,01): Esse valor alto (acima de 10) confirma uma associação extremamente forte entre os termos, sugerindo que é uma tradução consolidada.
5) Keyness (205,79): O número elevado indica que essa combinação é altamente relevante no contexto regulatório de Macau.

A extração automática de termos pode ser revisada para identificar e extrair o contexto definitório em dois idiomas. Esse processo é crucial para criar glossários mais detalhados, que são úteis para serem usados diretamente em ferramentas de tradução como Matecat, MemoQ ou Wordfast. Esses glossários podem ser exportados no formato TBX (Term Base Exchange), facilitando a integração com essas ferramentas. A exploração de nosso corpus paralelo alinhado com AntPconc permite validar a terminologia selecionada pelo OneClick Terms. A expressão de busca "相關 物品" informa que temos de fato 168 ocorrências no corpus chinês e retorna as seguintes informações (escolha de trechos curtos para melhorar a leitura e compreensão.):

Tabela 10: Exemplo de terminologia extraída com o OneClick Terms e validada por meio do AntPConc

Fonte: Corpus_Leis_zh-tkn.txt	**Fonte: Corpus_Leis_pt.txt**
武器 及 相關 物品	Armas e coisas conexas
禁用 的 武器 及 相關 物品	Armas e coisas conexas proibidas
受管 控 武器 及 相關 物品	Armas e coisas conexas controladas
扣押 武器 及 相關 物品	Apreensão de armas e coisas conexas

Fonte: Autor (2025)

A tradução da TAN de Baidu "itens relacionados", não consta no corpus em português. As expressões "itens relacionados" e "coisas conexas" têm significados semelhantes, mas não são exatamente sinônimos.

A consulta de corpus do Antconc permite identificar mais informações. A expressão de busca "武器 及 相關 物品", com 130 ocorrências, encontra-se apenas na Lei n.º 12/2024 sobre o "Regime jurídico do controlo de armas e coisas conexas". Da mesma forma, "promotor(es) de jogo" (博彩中介), com 179 ocorrências, encontra-se nas duas leis do corpus, relacionadas à categoria de "Jogos

de Fortuna e Azar e Turismo: Lei nº 16/2022: Regime da exploração de jogos de fortuna ou azar em casino e a Lei nº 7/2022: Alteração ao Regime jurídico da exploração de jogos.

Tabela 11: Exemplo de terminologia extraída com o OneClick Terms e validada por meio do AntPConc

Fonte: Corpus_Leis_zh-tkn.txt	Fonte: Corpus_Leis_pt.txt
博彩 中介	Promotores de jogo
博彩 中介 佣金 的 税項	Imposto sobre as comissões pagas a promotores de jogo
博彩 中介 准照 的 強制性	Obrigatoriedade de licença de promotor de jogo
博彩 中介 合同	Contrato de promoção de jogos

Fonte: Autor (2025)

A frequência e a variedade de contextos em que o termo aparece sugerem que "promotor(es) de jogo" é um conceito central nas leis que regulamentam os jogos de fortuna e azar. Um promotor de jogo (博彩中介) é um profissional que atua como intermediário no setor de casinos e jogos de fortuna e azar (娛樂場 幸運 博彩, 65 ocorrências), principalmente em jurisdições como Macau. Segundo a legislação do corpus, eles desempenham um papel crucial no ecossistema dos casinos, recrutando jogadores VIP de alto valor e organizando sua logística completa incluindo viagens, hospedagem e entretenimento personalizado. Esta abordagem baseada em corpus oferece ao tradutor não apenas equivalentes isolados, mas também contextos de uso autênticos, essenciais para capturar nuances culturais e jurídicas que simples dicionários não contemplam, resultando em traduções que respeitam tanto a precisão terminológica quanto a naturalidade discursiva.

3.3 Exploração do corpus chinês com Sketch Engine: 3-grams e 4-grams

Esta exploração objetiva analisar os 100 principais 3-grams e 4-grams jurídicos em chinês de Macau, extraídos via Sketch Engine com base em sua frequência no corpus (**0**, p. 15) e organizados em uma planilha Excel[6]. O objetivo dessa exploração é verificar, a partir desses termos mais recorrentes, as observações de Leong (2012) sobre a influência do português nos textos jurídicos macaenses, bem como sua relação com a Lei Básica de Macau.

Cada termo foi categorizado semanticamente com apoio do LLM chinês, Deepseek (2023), (coluna "Categoria", Tabela 12, abaixo) e avaliado quanto a: (1) probabilidade de ser um *calque* do português e (2) impacto na compreensão. Para isso, compararam-se sistematicamente as traduções fornecidas por LLMs (Deepseek, Mistral, Claude, GPT) e TANs (Baidu, Lara Translate, Google Translate). A análise focou especialmente nos padrões de hibridismo linguístico e na permeabilidade de estruturas lusófonas na terminologia jurídica local. Os resultados demonstraram a coexistência de padrões de *calques* portugueses, confirmando em parte as hipóteses de Leong (2012) sobre a dupla influência jurídico-linguística portuguesa em Macau. A metodologia combinou análise quantitativa (frequência de n-grams) com avaliação qualitativa das traduções, oferecendo uma visão estratificada do fenômeno jurídico-linguístico.

[6] Os scripts e os corpora criados nesta pesquisa encontram-se disponíveis no repositório de acesso aberto Figshare (Miroir, 2025).

3.3.1 Distribuição por categoria jurídica

Tabela 12: Distribuição por categoria jurídica

Categoria	Quantidade	% do Total
Referências jurídicas	18	18%
Estrutura jurídica	6	6%
Textos jurídicos	10	10%
Disposições	12	12%
Erros	16	16%
Entidade territorial	7	7%
Processos	4	4%
Infrações	3	3%
Outros (moeda, jurisdição, etc.)	24	24%
Total	100	100%

Fonte: Autor (2025)

A análise da distribuição por categoria dos 100 termos jurídicos em chinês revela que as "Referências jurídicas" constituem a categoria mais expressiva (18%), seguida pelas "Disposições" (12%) e "Textos jurídicos" (10%), refletindo a natureza técnica e intertextual dos documentos jurídicos. Os "Erros" (16%) aparecem como segunda categoria mais frequente, majoritariamente decorrentes de problemas de tokenização. As categorias mais específicas como "Processos" (4%) e "Infrações" (3%) apresentam menor representatividade, enquanto o grupo "Outros" (24%) engloba termos diversos como moeda e jurisdição, demonstrando a variedade terminológica do universo jurídico de Macau. Essa distribuição evidencia a predominância de termos de remissão e disposição normativa, característicos da linguagem jurídica, com uma significativa ocorrência de desafios técnicos na interpretação automática.

3.3.2 Classificação como *calque* do português

Tabela 13: Classificação como *calque* do português

Classificação	Quantidade	% Válido*
Sim (*Calque*)	56	66,7%
Não	40	47,6%
Parcial	4	4,8%
Subtotal (Válidos)	84	84,0%
N/A (Erros/Fragmentos)	16	–

** Nota: % calculada sobre 84 entradas válidas (excluindo N/A).*

Fonte: Autor (2025)

A análise dos termos jurídicos em chinês de Macau revela uma significativa influência da língua portuguesa, com 66,7% dos casos (56 termos) classificados como *calques* linguísticos - incorporando estruturas ou expressões diretamente derivadas do português. Os 40 termos restantes (47,6%) não apresentam essa influência, enquanto 4 casos (4,8%) demonstram adaptação parcial. Esta predominância de *calques* reflete claramente o legado do sistema jurídico português em Macau,

particularmente em documentos legais e terminologia oficial. Os 16 casos excluídos da análise (N/A) correspondem a fragmentos ou erros de tokenização que não permitiram classificação adequada, mas não afetam a conclusão principal sobre a forte presença de estruturas portuguesas no vocabulário jurídico local.

3.3.3 Impacto dos *calques* na compreensão

Tabela 14: Impacto dos *calques* na compreensão

Nível de impacto	Quantidade	% Válido*
Alto	24	27,30%
Moderado	32	36,40%
Leve	8	9,10%
Neutro/Nenhum	24	27,30%
Subtotal (Válidos)	88	88%
N/A	12	–

**Nota: % calculada sobre 88 entradas válidas (excluindo N/A).*

Fonte: Autor (2025)

A análise do impacto na compreensão dos termos jurídicos demonstra que 63,7% das expressões (56 termos entre moderados e altos) apresentam dificuldades significativas de interpretação, sendo que 27,3% (24 termos) possuem impacto alto por sua complexidade ou ambiguidade. Outros 27,3% (24 termos) são neutros e não apresentam obstáculos à compreensão, enquanto apenas 8 termos (9,1%) têm impacto leve. Esses resultados evidenciam os desafios na interpretação da linguagem jurídica de Macau, onde mais da metade dos termos exige conhecimento especializado ou contextualização adequada para seu correto entendimento. Os 12 casos não classificados (N/A) correspondem a fragmentos ou erros de tokenização que não permitiram avaliação precisa.

3.3.4 Frequência de termos mais comuns

Tabela 15: Frequência de termos mais comuns

Termo em Chinês	Tradução mais frequente	Ocorrências	Freq. corpus
澳門 特別 行政區	"Região Administrativa Especial de Macau"	8	1259
款 所 指	"referido no parágrafo"	6	734
條 的 規定	"disposições do artigo/parágrafo"	5	295
文本 所 表述	"expresso no texto"	4	452

Fonte: Autor (2025)

Os termos jurídicos mais frequentes na análise apresentam traduções altamente padronizadas, demonstrando consistência terminológica nos documentos jurídicos de Macau. A tradução da expressão "澳門 特別 行政區" (Região Administrativa Especial de Macau) aparece 8 vezes, confirmando sua natureza de denominação oficial obrigatória. Da mesma forma, "款 所 指" (referido no parágrafo) com 6 ocorrências e "條 的 規定" (disposições do artigo/parágrafo) com 5 ocorrências revelam a adoção sistemática de fórmulas jurídicas portuguesas. O termo "文本 所 表

述” (expresso no texto), presente em 4 casos, completa esse conjunto de expressões padronizadas que formam o núcleo da linguagem jurídica institucional na região, evidenciando a forte influência do português na redação de documentos oficiais.

3.3.5 Análise de Erros

Tabela 16: Análise de Erros

Tipo de Erro	Quantidade	Exemplo
Tokenização incompleta	12	“澳門 特別 行政” (faltando “區”)
Termos fragmentados	4	“葡 文 文”

Fonte: Autor (2025)

A análise identificou 16 casos de erros no processamento de termos jurídicos, sendo 75% (12 ocorrências) classificados como tokenização incompleta, onde faltam caracteres essenciais, como no exemplo “澳門 特別 行政” que omite o caractere final “區”. Os 25% restantes (4 casos) correspondem a termos fragmentados, como “葡 文 文”, que apresentam repetições ou estruturas incompletas. Estes erros totalizam 16% do corpus de n-grams analisado e comprometem a precisão das traduções automáticas e destacam a necessidade de processos de pré-processamento mais robustos para garantir a qualidade na interpretação de documentos jurídicos em chinês de Macau.

3.3.6 Consistência nas traduções

Tabela 17: Consistência nas traduções

Métrica	Resultado
Concordância em termos oficiais (ex.: RAEM)	92%
Divergências em termos técnicos	8%

Fonte: Autor (2025)

Essa análise comparou traduções automáticas de termos jurídicos entre modelos de linguagem (LLMs - Deepseek, Mistral, Claude, GPT) e sistemas de tradução automática (TANs - Baidu, Lara Translate, Google Tradutor). Foram analisados dois grupos de termos, como: oficiais (nomes padronizados) e técnicos (expressões com múltiplas interpretações). Foram contabilizados 25 termos oficiais com traduções consolidadas e expressões técnicas com possíveis variações. Para cada termo, verificou-se a uniformidade entre as ferramentas. O cálculo considerou apenas traduções idênticas, excluindo fragmentos e erros. Os termos oficiais apresentaram 92% de concordância (23/25 casos), como “澳門特別行政區” traduzido unanimemente como “Região Administrativa Especial de Macau”. Já os termos técnicos mostraram 8% de divergência, com variações como “processo executivo” versus “procedimento de execução”.

As ferramentas (LLMs e TANs) demonstraram excelente consistência em traduções institucionais, seguindo convenções estabelecidas. Contudo, expressões jurídicas específicas apresentaram pequenas variações, principalmente entre LLMs (mais descritivos) e TANs (mais concisos). Isso reforça a necessidade de revisão humana para termos técnicos complexos.

3.3.7 Principais observações

Essa breve análise revela uma dominância de *calques* linguísticos, com 66,7% dos termos válidos reproduzindo estruturas sintáticas do português no texto chinês. Erros comuns, como fragmentação lexical e tokenizações incompletas, representam cerca de 16% dos casos analisados, comprometendo a fluidez e a precisão terminológica. Observou-se ainda um alto grau de consistência e padronização em termos oficiais (92%), contrastando com uma maior variação na tradução de termos técnicos.

A análise comparativa entre LLMs (modelos de linguagem) e TANs (sistemas de tradução automática) revela diferenças significativas em sua aplicação para textos jurídicos. Os LLMs como Deepseek (2023) e Claude (2024) destacam-se por oferecer explicações contextuais mais ricas, embora com tendência a redundâncias, como transformar simplesmente "referido" em "o que é referido". Por outro lado, TANs como Google Tradutor e Baidu apresentam maior precisão em traduções diretas e padronizadas, especialmente em termos oficiais consagrados. Para documentos jurídicos de alta qualidade, recomenda-se uma abordagem híbrida: utilizar TANs como base para garantir fidelidade terminológica, complementando com LLMs para esclarecer nuances contextuais, seguido sempre por revisão humana para eliminar redundâncias e garantir precisão absoluta, particularmente em termos com alto impacto na interpretação jurídica.

3.4 Exploração do corpus chinês e português

Esta exploração tem como objetivo evidenciar, em nosso corpus paralelo, padrões sintáticos que caracterizam influências do português para o texto jurídico chinês. As expressões de busca foram elaboradas com base nas características do corpus, com o apoio do LLM chinês DeepSeek.

3.4.1 Padrões de ordem sintática incomum de uso de estrutura portuguesa

Expressão de busca: "從事 * 的"

Objetivo: Encontrar construções do tipo 從事 [atividade] 的 "que exercem [atividade]" (Cláusula relativa) ou "para o exercício de [atividade]" (Nominalização).

44 ocorrências do padrão encontradas.

Tabela 18: Exemplos de padrões gramaticais extraídos e validados por meio do AntPConc

Fonte: Corpus_Leis_zh-tkn.txt	Fonte: Corpus_Leis_pt.txt
小販 及 船舶 經紀人 在 海事 管 轄範 圍內 **從事 工作 的** 准照 ：	Licença para vendilhões e correctores **exercerem os seus misteres** na área de jurisdição marítima:
（ 一 ） **從事 業務 的** 准入 制度 ；	1) O regime de acesso ao exercício de actividade;
（ 三 ） 已 履行 所 **從事 活動 的** 稅務 義務 ；	3) Ter cumprido as obrigações fiscais inerentes **à actividade que exerce**;

Fonte: Autor (2025)

A estrutura "從事 * 的" revela a influência do português, traduzindo tanto cláusulas relativas ("que exercem") como nominalizações ("para o exercício de"). A ordem sintática confirma a adaptação de padrões gramaticais portugueses em contextos jurídicos de Macau.

3.4.2 Preposições e conectores portugueses

Expressão de busca: "為適 用 *"

Objetivo: Localizar a construção do tipo 為適用 [algo] calcada do português "para efeitos de [algo]". 227 ocorrências do padrão encontradas.

Tabela 19: Exemplos de padrões gramaticais extraídos e validados por meio do AntPConc

Fonte: Corpus_Leis_zh-tkn.txt	Fonte: Corpus_Leis_pt.txt
為適 用 本 法律， 下列 用語 的 含義 為：	Para efeitos da presente lei, entende-se por:
為適 用 本 法律 的 規定， 下列 用語 的 含義 為：	Para efeitos do disposto na presente lei, entende-se por:

Fonte: Autor (2025)

A expressão "為適用 *" é um empréstimo sintático do português "para efeitos de *", comum em textos jurídicos. Sua alta frequência, no corpus, em chinês jurídico de Macau mostra a influência portuguesa, adaptando-se gramaticalmente ao chinês. Isso evidencia um caso de empréstimo sintático consolidado, onde a expressão portuguesa foi lexicalizada no discurso jurídico bilíngue.

3.4.3 Estruturas passivas híbridas

Expressão de busca 1: "經 * 證明".

Objetivo: Encontrar a construção do tipo 經 [algo] 證明 calcada da estrutura passiva portuguesa "conforme comprovado por [algo]". 1 ocorrência do padrão encontrada.

Expressão de busca 2: "由 * 證明" (Voz passiva analítica) do tipo "é comprovado por *" / "fica certificado por *". 1 ocorrência do padrão encontrada.

Expressão de busca 3: "* 確認" para "* confirma" / "* certifica". 146 ocorrências do padrão encontradas.

Tabela 20: Exemplos de padrões gramaticais extraídos e validados por meio do AntPConc

Fonte: Corpus_Leis_zh-tkn.txt	Fonte: Corpus_Leis_pt.txt
法律 效力， 利害 關係人 不 遞交 治安 警察局 要求 其 提供 的 資料， 或 在 應 領取 的 期間 不 領取 准照 或 取得 受管 控 武器 或 裝置、 相關 彈藥 或 投射物 的 預先 許 可 的 憑證， 等同 放棄 准照 或 許可， 但 有 合理 且 **經 適當 證明** 的 理由 除外。	2. Salvo motivo justificado e **devidamente comprovado**, a falta de entrega dos elementos que o CPSP tenha solicitado ao interessado, ou o não levantamento da licença ou do título da autorização prévia de aquisição de armas ou dispositivos controlados ou das respectivas munições ou projécteis, nos prazos devidos, equivale, para todos os efeitos legais, à renúncia à autorização ou licença.
二、 就 義務 的 不 履行， 須 **由 受託人 證明** 非 因 其 過錯 所 造成。	2. **Incumbe ao fiduciário provar** que o incumprimento dos deveres não procede de culpa sua.
已 於獲 **政府 確認** 的 遊艇 俱樂部 登記：	Registadas nos clubes náuticos, **reconhecidos pelo Governo**:
（ 五 ） 根據 第五 章 的 **規定 確認** 國家 秘密 事項 及 解除 保密 義務	5) **Confirmar as matérias** de segredo de Estado e levantar o dever de sigilo nos termos do disposto no capítulo V;

Fonte: Autor (2025)

A estrutura "* 確認" constitui a forma idiomática predominante no corpus para expressar confirmações, seguindo a sintaxe ativa (sujeito-verbo-objeto) natural desta língua. Sua alta frequência no corpus (146 ocorrências) em contraste com as raras construções passivas calcadas do português comprova, no caso, esta preferência linguística, oferecendo orientação valiosa para tradutores produzirem versões mais naturais ao converter passivas portuguesas em ativas chinesas.

Esta breve pesquisa, realizada através do AntPconc com base em expressões de busca sugeridas pelo Deepseek para identificar *calques* português no corpus, demonstra que o chinês jurídico de Macau incorpora estruturas do português de forma estratégica, combinando empréstimos sintáticos com adaptações locais.

4. Conclusão

Neste artigo, apresentamos uma análise abrangente e detalhada do bilinguismo jurídico em Macau, destacando a singularidade desse fenômeno no contexto do princípio "um país, dois sistemas". A contextualização histórica e jurídica da Região Administrativa Especial de Macau (RAEM) foi essencial para compreender a coexistência do chinês e do português em um sistema jurídico híbrido. Essa base conceitual permitiu explorar a hierarquia linguística entre o chinês, falado pela maioria da população, e o português, que desempenha um papel técnico e suplementar, especialmente no âmbito jurídico.

Ao longo do trabalho, identificamos os desafios de tradução jurídica, como traduções literais e a escassez de profissionais bilíngues qualificados, e propusemos soluções práticas, como a inversão do fluxo de tradução e o investimento na formação de tradutores jurídicos. Essa abordagem crítica e propositiva é um diferencial que enriquece o debate sobre a questão do bilinguismo em Macau.

A metodologia empregada neste estudo apresenta uma abordagem inovadora. A compilação e análise dos corpora paralelos de leis em português e chinês foram descritas de forma clara e detalhada, com um foco na automatização de processos e no uso de ferramentas avançadas como Bertalign e spaCy. O alinhamento automatizado de textos jurídicos, em particular, demonstrou a complexidade e a precisão do método adotado. A utilização de ferramentas especializadas, como Antconc, AntPconc e Sketch Engine, permitiu análises profundas e comparativas dos corpora, fortalecendo o embasamento técnico da pesquisa.

A análise de *calques* linguísticos constituiu um aspecto central deste artigo. A influência do português nos textos jurídicos em chinês foi bem fundamentada, com exemplos concretos que comprovam as afirmações de Leong (2012) sobre a presença de estruturas portuguesas no chinês jurídico de Macau. O apoio de um LLM chinês (Deepseek) para as pesquisas sobre a língua chinesa foi um recurso valioso, garantindo a precisão e a relevância das análises. Os dados obtidos são ricos e exploráveis, oferecendo uma base sólida para futuras investigações.

A exploração dos corpora em chinês e português foi extensa e abrangente, incluindo análises de n-grams, padrões sintáticos e terminologia jurídica. A comparação entre LLMs (modelos de linguagem) e TANs (sistemas de tradução automática) para tradução jurídica foi relevante e bem fundamentada, oferecendo *insights* sobre a precisão e a consistência das traduções. Abordamos de forma crítica os erros de tokenização e os desafios de tradução, propondo soluções práticas que podem ser aplicadas em contextos reais.

Este artigo contribui significativamente para o campo da linguística de corpus, especialmente em contextos jurídicos multilíngues. A pertinência do trabalho de compilação, a variedade dos exemplos de pesquisa e a profundidade das análises fazem deste estudo uma referência instigante para pesquisadores, tradutores e profissionais do direito que lidam com questões de bilinguismo jurídico. A combinação de rigor metodológico, a abordagem metodológica e a aplicabilidade em contextos reais tornam este trabalho uma contribuição significativa para o avanço do conhecimento nessa área. Vale ressaltar que essa pesquisa teria sua qualidade significativamente ampliada com a validação dos resultados por falantes nativos de chinês, residentes em Macau e familiarizados com a linguagem jurídica específica.

Referências

Anthony, L. (2024a). *Antconc* (Version 4.0) [Software].

Anthony, L. (2024b). *AntPconc* (Version 1.2.1) [Software].

Anthony, L. (2024c). TagAnt [Software]

bfsujason. (2022). Bertalign (versão 0.1.0) [Software]. GitHub.

Boletim Oficial (B.O.) de Macau. (1988). *Declaração Conjunta do Governo da República Portuguesa e do Governo da República Popular da China sobre a Questão de Macau [com Anexos I e II]*. https://bo.io.gov.mo

Boletim Oficial (B.O.) de Macau. (1993). *Lei Básica da Região Administrativa Especial de Macau da República Popular da China, de 31 de março de 1993, Promulgada pelo Decreto n.º 3 do Presidente da República Popular da China*. https://bo.io.gov.mo

Cheng, L., & Sun, Y. (2021). Terminology translation in socio-legal contexts: A corpus-based exploration. In S. Li & W. Hope (Eds.), *Terminology Translation in Chinese Contexts: Theory and Practice* (pp. 27–39). Routledge.

Claude AI. (2024). *Claude AI* [Software].

Deepseek. (2023). *Deepseek* [Software].

Gao, Z.-M. (2021). Automatically compiling bilingual legal glossaries based on Chinese-English parallel corpora. In S. Li & W. Hope (Eds.), *Terminology Translation in Chinese Contexts: Theory and Practice* (p. 164–179). Routledge. http://doi.org/10.4324/9781003006688-14

Imprensa Oficial (IO) do Governo da Região Administrativa Especial de Macau. (2025). *Página inicial*. io.gov.mo.

Lefer, M.-A. (2020). Parallel corpora. In M. Paquot & S. T. Gries (Eds.), *A Practical Handbook of Corpus Linguistics* (p. 257–282). Springer Nature.

Leong, S. M. (2012). Divergências linguísticas e interpretação correcta da Lei Básica. *Revista de Estudos de "Um País, Dois Sistemas"*, *4*, 183–193.

Liu, L., & Zhu, M. (2022). Bertalign: Improved word embedding-based sentence alignment for Chinese–English parallel corpora of literary texts. *Digital Scholarship in the Humanities*, *38*(2), 621–634. https://doi.org/10.1093/llc/fqac089

Miroir, J.-C. (2024a). Compilação e exploração de material de apoio à tradução de textos jurídicos normativos: o caso da versão do português para o francês (*AntPconc*). In F. C. C. L. Arraes,

A. R. de Oliveira Harden & C. Roscoe-Bessa (Eds.), *Tradução em contextos específicos: conhecimentos e habilidades* (pp. 13–49). Pontes Editores.

Miroir, J.-C. (2024b). Processamento de linguagem natural multilíngue com spaCy e análises avançadas de corpora anotados com Antconc (versão 4). In *Encontro de Linguística de Corpus & Escola Brasileira de Linguística Computacional ELC/EBRALC, Universidade de Brasília, 21–24 de Outubro, 2024*. [Workshop handout]. Departamento de Línguas Estrangeiras e Tradução, Instituto de Letras, Universidade de Brasília. https://doi.org/10.13140/RG.2.2.24082.67520

Miroir, J.-C. (2025). *Tradução jurídica em contexto (TraJeC): Bilinguismo jurídico chinês-português em Macau* [Data set]. Figshare Datacite. https://figshare.com/projects/TraJeC

Newman, J., & Cox, C. (2020). Corpus Annotation. In M. Paquot & S. T. Gries (Eds.), *A Practical Handbook of Corpus Linguistics* (pp. 24–48). Springer Nature Switzerland AG. https://doi.org/10.1007/978-3-030-46216-1_2

Paquot, M., & Gries, S. T. (Eds.). (2020). *A Practical Handbook of Corpus Linguistics*. Springer Nature Switzerland AG.

Sardinha, T. B. (2000). Lingüística de Corpus: histórico e problemática. *DELTA*, *16*(2). https://doi.org/10.1590/S0102-44502000000200005

Notas editoriais

Contribuição de autoria

Concepção e elaboração do manuscrito: J.-C. L. Miroir
Coleta de dados: J.-C. L. Miroir
Análise de dados: J.-C. L. Miroir
Discussão dos resultados: J.-C. L. Miroir
Escrita: revisão e aprovação: J.-C. L. Miroir

Conjunto de dados de pesquisa
Os dados desta pesquisa estão disponíveis no repositório de dados Figshare (Miroir, 2025).

Financiamento
Não se aplica.

Consentimento de uso de imagem
Não se aplica.

Aprovação de comitê de ética em pesquisa
Não se aplica.

Conflito de interesses
Não se aplica.

Declaração de disponibilidade dos dados da pesquisa
Os dados desta pesquisa, que não estão expressos neste trabalho, poderão ser disponibilizados pelo(s) autor(es) mediante solicitação.

Publisher
Cadernos de Tradução é uma publicação do Programa de Pós-Graduação em Estudos da Tradução, da Universidade Federal de Santa Catarina. A revista *Cadernos de Tradução* é hospedada pelo Portal de Periódicos UFSC. As ideias expressadas neste artigo são de responsabilidade de seus autores, não representando, necessariamente, a opinião dos editores ou da universidade.

Histórico
Recebido em: 30-04-2025
Aprovado em: 24-06-2025
Revisado em: 30-07-2025
Publicado em: 09-2025