

KPEEC: Corpus paralelo coreano-português-inglês de esports

KPEEC: Korean-Portuguese-English *esports* parallel corpus

Vítor Amaral Córdoba¹
Thiago Blanch Pires²

Resumo: Este trabalho explora um *corpus* paralelo coreano-português-inglês com enfoque no domínio dos *esports*, mais especificamente no jogo eletrônico intitulado *League of Legends*. O *corpus* foi montado a partir das documentações oficiais dos jogos e de transcrições de falas de jogadores profissionais da liga coreana durante jogos oficiais e transmissões de partidas *online*, a fim de incluir tanto a linguagem oficial presente nos arquivos do jogo quanto a linguagem utilizada pela comunidade de jogadores. A coleta de dados da língua coreana e da mesma com tradução somente para o inglês foi realizada pelos autores com tradução para o português. A estrutura do *corpus* permite o estudo tradutório e comparativo entre as línguas, uma vez que há uma divisão em três *subcorpora* que incluem o uso da língua em contextos diversos. Os resultados revelam similaridades entre a terminologia utilizada entre as três línguas, frequentemente emergidas de sua origem no inglês. Os resultados também demonstram uma variação entre a denominação oficial dos termos listada na documentação dos jogos e o uso dos mesmos pela comunidade de jogadores profissionais coreanos.

Palavras-chave: *Esports*. Linguística de *Corpus*. Língua coreana. Terminologia. Tradução.

Abstract: This study aims to explore a Korean-Portuguese-English parallel corpus focusing on the *esports* domain, specifically the electronic game entitled “League of Legends”. The corpus was compiled from official game documentation and transcriptions of professional players’ speeches in the Korean league during official matches and online broadcasts, to include both the official language present in game files and the language authentically used by the gaming community. For the collection of data in Korean and its translation into English, translation into Portuguese was performed by us. The corpus structure allows for translational and comparative studies between the languages, as it is divided into three subcorpora that reflect the use of language in different contexts. Similarities were found in the terminology used across the three languages, often derived from their English origins. The results also show noticeable variation was also observed between the official designation of terms listed in the game documentation and the usage of these terms by the Korean professional gaming community.

Keywords: *Esports*. *Corpus* Linguistics. Korean language. Terminology. Translation.

Introdução

Os esportes eletrônicos, amplamente conhecidos como *esports*, surgiram como uma forma de entretenimento e competição de rápido crescimento. Essa indústria diversificada e moderna, que abrange uma variedade de *videogames*, conquistou uma audiência global relevante. Com eventos lotados, transmissões ao vivo que atraem milhões de espectadores

¹ Universidade de Brasília, Instituto de Letras, Departamento de Línguas Estrangeiras e Tradução, Brasília, Distrito Federal, Brasil. Endereço eletrônico: vitorcordoba1@gmail.com.

² Universidade de Brasília, Instituto de Letras, Departamento de Línguas Estrangeiras e Tradução, Programa de Pós-Graduação em Linguística, Brasília, Distrito Federal, Brasil. Endereço eletrônico: pirestb@unb.br.

e atletas profissionais ganhando reconhecimento internacional, os *esports* consolidaram sua posição como uma parte vital da cultura contemporânea e são responsáveis por fomentar a indústria dos *videogames*, a principal na categoria entretenimento³.

No epicentro desse fenômeno, encontra-se a Coreia do Sul, um país que está entre um dos maiores destaques mundiais dos *esports*⁴. Com uma rica história de competições de *videogames* e um rápido desenvolvimento tecnológico, a Coreia do Sul tornou-se um bastião inegável dos *esports*, influenciando a indústria global de sobremaneira. No entanto, apesar do sucesso dos *esports* sul-coreanos, uma barreira permanece: a barreira linguística.

Embora os sul-coreanos tenham uma influência notável nos *esports*, o acesso a conteúdo relevante e informações detalhadas sobre esse mundo muitas vezes se mostra inacessível para aqueles que não falam a língua coreana. Os termos específicos e as abreviações utilizadas pelas comunidades de *esports* sul-coreanas frequentemente constituem uma linguagem de especialidade, obscurecendo ainda mais o entendimento para aqueles que não estão familiarizados com ela.

O objetivo geral desta investigação é explorar um *corpus* paralelo coreano-português-inglês de *esports*. O corpus inclui os termos usados pelos falantes coreanos em jogos *online*, com destaque para o jogo *League of Legends*, devido ao seu impacto e influência no meio dos *esports*. Isso permitirá uma compreensão mais profunda e acessível desse cenário de *esports* sul-coreanos para a audiência brasileira, além de servir como um *corpus* de referência em uma área de conteúdo escasso.

As informações diretamente relacionadas à criação do *corpus* serão introduzidas na terceira seção, qual seja, a metodologia. Após essa etapa, descreveremos a metodologia utilizada para compilar um *corpus* paralelo coreano-português-inglês, que incluirá os termos utilizados pelos falantes coreanos em jogos *online* na terceira seção.

Por fim, apresentaremos os resultados da pesquisa na quarta seção e discutiremos possíveis aplicações do corpus e delinearemos planos futuros para a continuidade deste trabalho, destacando sua contribuição para tornar o conteúdo dos *esports* sul-coreanos mais acessível e compreensível para a audiência brasileira na quinta seção.

Fundamentação teórica

Estudos de corpora e tradução

Diferentes estudos podem apresentar divergências na definição e categorização dos diferentes tipos de *corpora*. Em nosso estudo, busca-se a criação de um *corpus* paralelo,

³ Estudos da *TechNETT Immersive*.

⁴ Segundo relatório da Newzoo. Disponível em: <<https://newzoo.com/resources/rankings/top-10-countries-by-game-revenues>>. Acesso em: 15 mai. 2025.

conforme a definição de teóricos relevantes da área da Linguística de *Corpus* e Estudos da Tradução (Baker, 1993; Olohan, 2004; Johansson, 2007; Dayrell, 2015). Um *corpus* paralelo é uma coleção de textos ou dados linguísticos em duas ou mais línguas que estão alinhados de forma que seja possível identificar correspondências entre segmentos linguísticos equivalentes. Ou seja, é uma ferramenta que permite a análise de textos em diferentes idiomas em paralelo, possibilitando a identificação de traduções e equivalências linguísticas (Baker, 1993).

Apesar de muitas vezes os *corpora* paralelos serem bilíngues, eles também podem ser multilíngues, incluindo traduções para mais de um idioma a partir de um mesmo texto-fonte. O *corpus* desenvolvido neste trabalho, por incluir os idiomas inglês, coreano e português, é, portanto, um *corpus* paralelo multilíngue. Há, porém, uma limitação: a ausência de textos originais em português com tradução para os outros idiomas. Dessa forma, o *corpus* sofre no quesito de direcionalidade de suas línguas – tema tratado ao longo de nosso estudo.

Diferentemente de corpora como o COMPARA, composto por textos originais do inglês traduzidos para o português e de textos originais do português traduzidos para o inglês, ou seja, um *corpus* paralelo bidirecional (Dayrell, 2015, p. 93), o *corpus* elaborado neste estudo é parcialmente “unidirecional”, no sentido de que não é possível realizar a análise do estudo tradutório do português como língua-fonte. Isso se deve à poucas traduções do tema de textos originários do português para o coreano ou para o inglês. Dessa forma, apenas textos-fonte em coreano e em inglês foram utilizados para sua confecção.

A possibilidade de utilização de um *corpus* paralelo é diretamente ligada à disponibilidade de textos e de traduções. Por esse motivo, textos oficiais da União Europeia são frequentemente utilizados nessa modalidade (Olohan, 2004).

Johansson (2007) afirma que por meio dos *corpora* multilíngues podemos ver como línguas se assimilam e se diferem ao compará-las em escala maior, podendo possivelmente encontrar características sobre a língua geral como forma de comunicação humana. Porém, é também por meio dos *corpora* que podemos realizar inúmeras avaliações e trabalhar com os textos por meio de diferentes abordagens.

A principal abordagem retratada pelo autor é a da análise contrastiva (AC), que é a “comparação sistemática de duas ou mais línguas, com o objetivo de descrever suas semelhanças e diferenças” (Johansson, 2007, p.1, tradução nossa). Muitas vezes utilizada no meio acadêmico para o ensino de línguas, a AC necessita de estudos tradutórios e estudos comparativos para se coletar materiais suficientes para o propósito investigativo.

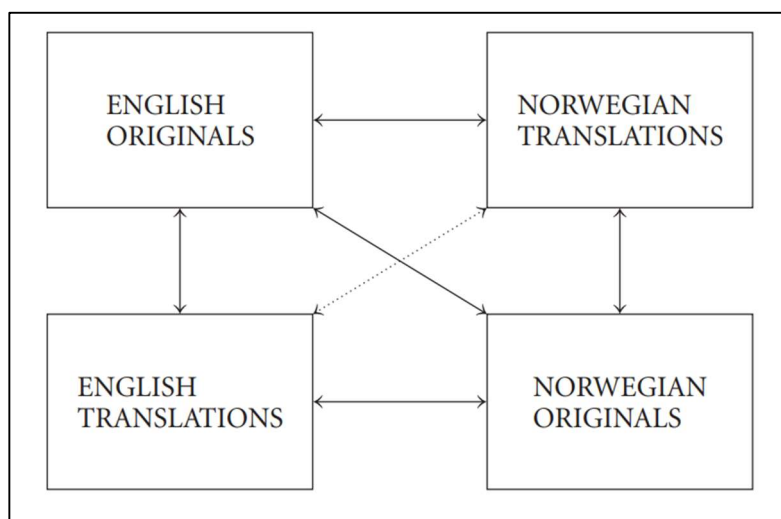
Uma vez que estudos tradutórios vão além da comparação sistemática entre línguas, mas ao mesmo tempo são limitados pela disponibilidade das traduções, pode ser significativo compreender a sua bidirecionalidade a partir da AC, de forma que a tradução pode fornecer dados para a análise contrastiva, e a análise contrastiva, por sua vez, auxilia a tradução a

solucionar problemas encontrados no processo (Hoey; Houghton, 1998, *apud* Johansson, 2007). Para alcançar os objetivos adequados de análise linguística após a criação de um *corpus*, o modelo e a formatação do *corpus* paralelo devem ser bem estabelecidos, possibilitando o diálogo entre estudos tradutórios e estudos contrastivos.

Entendemos aqui *corpora* que seguem o modelo tradutório como aqueles que contêm textos originais com traduções para uma ou mais línguas; e *corpora* comparáveis como aqueles constituídos de textos originais em duas ou mais línguas interligadas por critérios como gênero, período de publicação e outras características (Johansson, 2007).

O *corpus* desenvolvido no presente trabalho toma como base os conceitos de *corpus* tradutório e comparável do *Corpus Paralelo Inglês-Norueguês* – em inglês, *English-Norwegian Parallel Corpus* (ENPC) –, esboçado por Johansson (2007). A maior diferença entre este e aquele é a ausência de bidirecionalidade⁵ entre os pares coreano-português pela falta de textos originais em português.

Figura 1 – Estruturação do *English-Norwegian Parallel Corpus* (ENPC)



Fonte: Johansson (2007).

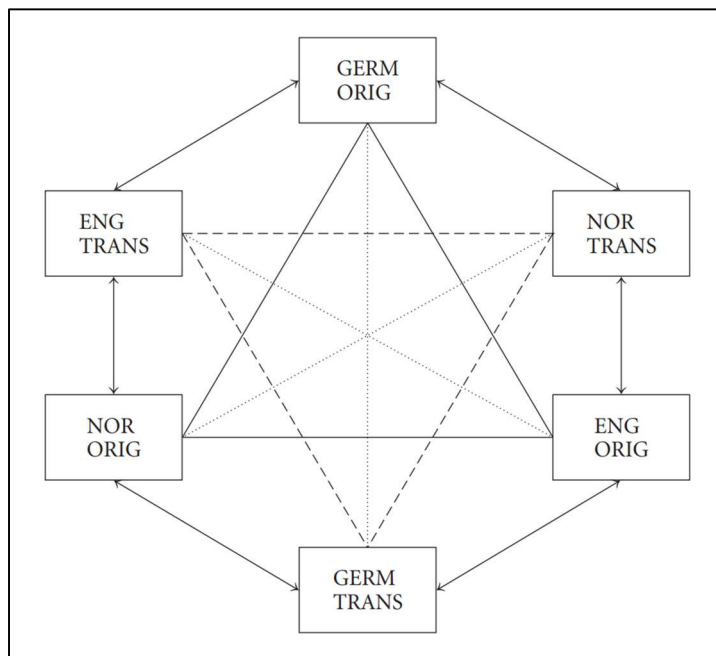
O *corpus* ENPC assume múltiplas funções. Pode ser utilizado como um *corpus* de tradução ao comparar textos originais e suas traduções entre os idiomas (setas horizontais); como *corpus* comparável entre textos originais (seta diagonal sólida e setas verticais) e entre textos traduzidos (seta diagonal pontilhada) (Johansson, 2007).

Uma extensão do *corpus* ENPC, o *Oslo Multilingual Corpus* (OMC) segue um modelo alternativo que é mais viável para a criação de um *corpus* plurilíngue – no caso específico

⁵ Entende-se a bidirecionalidade tanto de textos de uma língua A e suas traduções para uma língua B, quanto os textos originários dessa língua B traduzidos para a língua A (Johansson, 2007).

desse *corpus*, o inglês, o alemão e o norueguês. O modelo, denominado de “modelo diamante” por Johansson (2007), pode ser visto na Figura 2 a seguir.

Figura 2 – Estruturação do Oslo Multilingual Corpus (OMC)



Fonte: Johansson (2007).

Considerando as limitações da língua norueguesa similares às da língua coreana sob essa perspectiva de falta de recursos, o presente *corpus* busca solucionar o problema por meio de documentações oficiais dos jogos eletrônicos, uma vez que possuem adaptação e localização de recursos entre vários idiomas. A seleção de tais textos não necessariamente demonstra o vocabulário utilizado pela comunidade de jogadores e falantes da língua-alvo, mas permite que o *corpus* mantenha seu caráter tradutório e comparável.

Terminologia e corpus

O presente trabalho considera as orientações de Almeida e Correia (2008) como recurso-base para a criação de um *corpus* que permita uma pesquisa terminológica de maior qualidade. Em linha similar a Krieger e Finatto (2004), Almeida e Correia (2008) também defendem a importância de uma elaboração bem-conceituada do *corpus*, uma vez que o *corpus* compilado interfere diretamente com todas as etapas posteriores.

Sendo assim, a primeira informação relevante é elaborar um *corpus* em que seja possível realizar a extração de candidatos a termos, ou seja, “ítems léxicos que se comportam nos seus respectivos contextos como termos, mas cuja autenticidade será validada posteriormente” (Almeida; Correia, 2008, p. 74). É importante notar, em especial, duas

características de tal afirmação: (i) o contexto tem a sua devida importância e (ii) há a necessidade de uma validação de autenticidade terminológica.

A validação pode ser feita de diferentes formas, com uma recomendação maior àquela realizada de forma manual por especialistas do domínio. Apesar de ser o método recomendado, essa alternativa é muitas vezes substituída por uma comparação da lista de candidatos a um *corpus* de referência e contextualização de uso a partir dessas informações (Almeida; Correia, 2008, p. 75).

Referente à compilação do *corpus*, Almeida e Correia (2008) apontam seis etapas a serem seguidas para possibilitar pesquisas terminológicas. A primeira etapa na criação de um *corpus* terminológico é a "delimitação do domínio". Essa etapa visa restringir a área de estudo, evitando abordagens excessivamente abrangentes que podem se mostrar ineficazes. Ao delimitar um domínio, o pesquisador é capaz de selecionar subáreas de especialidade com contexto específico, o que impede diferenças significativas entre subáreas, ou necessidade de um trabalho extenso e complexo em um único *corpus* – tornando mais aconselhável a sua divisão em *corpora* para facilidade de análise após sua completude (Almeida; Correia, 2008, p. 77-79).

A segunda etapa é a "seleção das fontes de onde provirão os textos que deverão compor o *corpus*, bem como atenção aos requisitos recomendados e que já são consenso na Linguística de Corpus" (Almeida; Correia, 2008, p. 76). Nessa etapa, as autoras reforçam a necessidade de escolher fontes confiáveis para a construção do *corpus*. Essas fontes podem ser identificadas por meio de instituições, entidades, universidades e laboratórios envolvidos no domínio em questão, com preferência por fontes que disponham de material escrito, especialmente em formato digital. Também deverão ser levantadas questões como quantidade de termos, formatação do produto terminológico a ser gerado, público-alvo, variedade de fontes e outros fatores relevantes. A natureza do *corpus* a ser criado influenciará e será influenciada pela seleção dos textos que o comporão (Almeida; Correia, 2008, p. 79-82).

A terceira etapa, denominada "compilação (ou captura) dos textos", envolve a coleta e o armazenamento dos textos escolhidos em arquivos. Essa etapa pode ter grande variação a depender da fonte utilizada para a compilação do *corpus*, variando de uma simples cópia de arquivos de texto à necessidade de realizar um reconhecimento óptico de caracteres (OCR) para realizar a sua extração. Há vários métodos e casos de uso específicos para essa etapa (Almeida; Correia, 2008, p. 82-84).

A quarta etapa, "conversão de formatos e limpeza dos textos" (Almeida; Correia, 2008), requer a conversão, manual (recurso "copia-e-cola") ou automática, por meio de *softwares* disponíveis para a conversão de arquivos, como o XPDF, citado pelas autoras como alternativa gratuita. É recomendada a conversão de todos os arquivos para o formato "txt

(UTF-8)", dada a sua facilidade de manipulação. Nesse estágio, os textos também passam por um processo de limpeza e formatação, removendo elementos que fogem do formato textual, como imagens, gráficos e tabelas. (Almeida; Correia, 2008, p. 84).

Apesar da forte recomendação pelo uso do formato "txt" na conversão dos arquivos pelas autoras, o formato "json" é de extrema utilidade por sua estabilidade e formatação dinâmica para *corpus* multilíngue, como é o caso da formatação dos documentos utilizados pelo Instituto Nacional Sul-Coreano de Língua Coreana para a elaboração do *Korean-Foreign Language Parallel Corpus Project* (2023).

A quinta etapa, "nomeação dos arquivos", apesar de parecer simples, desempenha um papel importante na identificação e organização dos arquivos. Essa etapa é especialmente importante para *corpora* de maior volume, que contêm centenas ou milhares de arquivos de textos. Recomenda-se que os nomes sejam transparentes, a fim de facilitar a associação do terminólogo com o meio de divulgação, o gênero discursivo, a fonte e a data de publicação (Almeida; Correia, 2008, p. 85).

Por fim, a última etapa, "anotação dos textos", embora não seja obrigatória (demonstrada no livro como "se for o caso"), oferece benefícios significativos. A anotação, também denominada de etiquetagem, pode ser dividida em duas formas: a estrutural e a linguística. A anotação estrutural envolve a marcação de informações relevantes, tanto externas (como cabeçalhos) quanto internas (como a estrutura geral dos textos – por exemplo, capítulos, parágrafos, títulos e notas de rodapé). A anotação estrutural também engloba elementos como citações, abreviações, referências, datas e ênfases tipográficas (formatação textual), como negrito, itálico e sublinhado (Almeida; Correia, 2008, p. 86).

A anotação linguística pode ser realizada em diversos níveis: morfológico, morfossintático, semântico, dentre outros. Uma das características mais importantes desse tipo de anotação é a forma como ela é feita, dividida entre: automática, manual ou semiautomática. A anotação semiautomática é amplamente considerada a mais eficiente, pois possibilita uma revisão mais rápida do que a inteiramente manual e a geração de dados mais precisos do que a inteiramente automática (Almeida; Correia, 2008, p. 87).

Metodologia

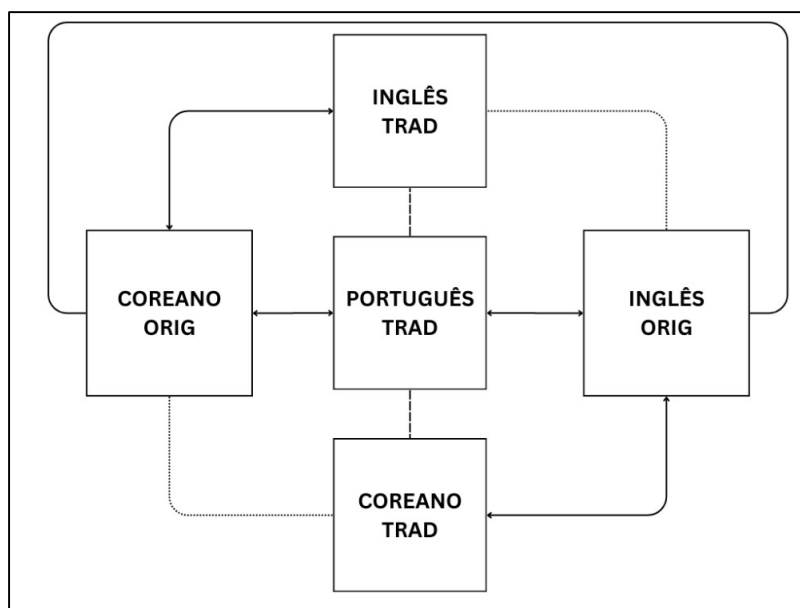
A metodologia tem como base as cinco primeiras etapas da criação de um *corpus* fundamentado para a análise terminológica delimitadas por Almeida e Correia (2008). A sexta etapa, das anotações, foi desconsiderada devido aos propósitos do presente trabalho. O formato geral dos textos que foram usados para compor o nosso *corpus* se encaixam, em sua maioria, na perspectiva do *corpus* paralelo multilíngue, com características mais específicas de ser de caráter tradutório, comparável e não bidirecional. Com o intuito de possibilitar maior completude ao *corpus*, também foram inclusos textos dos documentos oficiais de jogos em

coreano, inglês, e em português, a fim de tornar possível o estudo comparativo dos textos oficiais do jogo e de seu uso pela sua comunidade.

A elaboração do *corpus* é limitada essencialmente por dois fatores: tempo e disponibilidade de conteúdo. Conforme previamente comparado com o norueguês, não são abundantes os textos originais em coreano com equivalências em português e vice-versa. Tal limitação é um dos motivos para a inclusão da língua inglesa no *corpus*, visto que pode ser uma língua intermediária para apoiar as traduções e localizações de conteúdo em situações em que o português não é presente.

A Figura 3 toma como base a estrutura do *corpus* OMC e a adapta às necessidades de aplicação para o nosso *corpus*. Para melhor entender a figura, seguimos a mesma formatação de linhas e setas do OMC e indicamos textos originais como “ORIG” e traduções como “TRAD”. Essa estruturação assume uma função de *corpus* tradutório de textos ORIG e TRAD de línguas diferentes (cf. linhas sólidas com setas); utilização como *corpus* comparável entre TRAD de línguas diferentes (cf. linhas tracejadas); como *corpus* comparável entre ORIG e TRAD da mesma língua (cf. linhas sólidas pontilhadas) e como *corpus* comparável entre ORIG de línguas diferentes (cf. linhas sólidas sem setas). Assim como demonstrado por Johansson (2007) que o OMC não pôde ser construído exatamente de acordo com o seu modelo devido à falta de textos originais do norueguês, o *corpus* KPEEC não pôde ser completamente desenvolvido de acordo com o modelo OMC devido à falta de textos originais do português dentro do nosso escopo de uso.

Figura 3 – Estruturação do *corpus* paralelo coreano-português-inglês (KPEEC) no domínio de *esports*



Fonte: elaboração própria.

O nosso objetivo principal de pesquisa é o desenvolvimento do *Corpus* Paralelo Coreano-Português-Inglês de *Esports* (KPEEC), que permita o estudo tradutório, com especial enfoque no par coreano-português. Dessa forma, buscou-se incluir conteúdo suficiente em número de *tokens* e *types* para possibilitar este estudo. *Tokens* referem-se às unidades individuais que compõem o texto após a segmentação — representam o número total de ocorrências de todos os elementos linguísticos presentes no texto, incluindo repetições. *Types* correspondem ao conjunto de unidades únicas encontradas no *corpus*. Ou seja, cada palavra ou elemento distinto é contado apenas uma vez, independentemente do número de vezes em que aparece no texto⁶.

Para a coleta de textos, utilizamos fontes relevantes no domínio de *esports* para encontrar qual é a terminologia utilizada nesse universo e possivelmente encontrar comparações entre a linguagem oficial – descrita pela documentação do jogo – e a de uso real dos jogadores.

Cada proposta temática resultou em três *subcorpora* divididos internamente da seguinte maneira: 1) documentações de jogos; 2) transcrições da comunicação dos jogadores; 3) entrevistas realizadas com jogadores profissionais.

Para facilitar a identificação de cada *subcorpus* ao longo do texto, chamaremos o subcorpus 1 de *subcorpus* DJ, o *subcorpus* 2 de CJ e o *subcorpus* 3 de EJ.

Subcorpus 1: documentação de jogos (DJ)

Para a coleta de textos que constituiriam o *subcorpus* de documentações de jogos, utilizamos as bibliotecas API da *Riot Games* com o conteúdo que é base dos jogos de *League of Legends* e *Legends of Runeterra*. Uma API (*Application Programming Interface*) é uma interface que permite a comunicação entre diferentes sistemas e, muitas vezes, é utilizada para guardar dados e facilitar o seu acesso por meio de chaves API. A chave API, por sua vez, é um código exclusivo usado para autenticar e acessar os recursos de uma API.

A biblioteca *Data Dragon* disponibiliza a documentação dos jogos mais populares da desenvolvedora *Riot Games*: *League of Legends*, *Valorant*, *Teamfight Tactics* e *Legends of Runeterra*. Apesar disso, o jogo *Teamfight Tactics* não possui informações na API sobre seus arquivos do jogo, o que resultou na escolha apenas dos jogos *League of Legends* e *Legends of Runeterra*. *League of Legends* é um jogo de estratégia e batalhas em equipe em que o jogador controla um personagem (denominado “campeão”) e compra itens (como botas ou espadas) para tornar o seu campeão mais forte. *Legends of Runeterra* é um jogo de cartas estratégico baseado em turnos ambientado no mesmo universo, em que cada carta possui ações especiais.

⁶ O *corpus* não foi lematizado por limitações de tempo e recursos.

Os jogos escolhidos possuem documentação em português, inglês e coreano. Por ser uma empresa que tem como língua-base o inglês, supõe-se que esse *subcorpus* contemple textos originais em inglês com traduções para o português e coreano.

Para a coleta da API do *League of Legends* e do *Legends of Runeterra*, foi utilizada uma chave API em dois momentos: uma para extrair informações dos campeões, e outra para extrair informações dos itens. Tais informações são pontos cruciais dentro do jogo e contemplam conteúdos utilizados diretamente pelos jogadores. Os arquivos coletados estão em formato *.json*, e escolhemos apenas informações de texto relevantes – excluindo fatores como quantidade de vida de personagens, ou preços e valores de itens. O arquivo em formato *.json* é frequentemente utilizado para transmitir dados por uma rede, e sua capacidade de organização por meio de etiquetas em sua estrutura torna mais fácil o alinhamento de textos.

Os arquivos *.json* foram então divididos individualmente por campeão e item, no caso do *League of Legends*, e por carta, no caso do *Legends of Runeterra*. As *tags* presentes nos arquivos são idênticas em todos os idiomas, de tal forma que o mesmo texto sempre terá a mesma *tag* de identificação, o que facilita o alinhamento das entradas multilíngues do *corpus*.

Subcorpus 2: transcrições da comunicação dos jogadores (CJ)

Para o *subcorpus 2*, coletamos as transcrições da comunicação de jogadores profissionais em partidas oficiais e de falas dos jogadores em transmissões ao vivo por meio dos canais do *YouTube* de três das maiores organizações da liga coreana: *T1*, *Dplus KIA* e *Gen.G Esports*.

Todas as três organizações participam ativamente dos campeonatos mundiais anuais e coletivamente contam com milhões de seguidores em suas redes sociais. As transcrições em coreano foram extraídas com o alinhamento da legenda equivalente em inglês, e em seguida foram realizadas as traduções para o português por meio do tradutor automático *Google Cloud Translate*, com pós-edição manual para todas as entradas.

A maior dificuldade encontrada na obtenção dos dados foi a extração dos textos em coreano e inglês, uma vez que os vídeos do canal do *YouTube* da *T1* apresentam uma legenda implementada diretamente no vídeo – comumente denominada de *hardcoded subtitles*. O problema se repete parcialmente para os canais do *YouTube* da *Dplus KIA* e *Gen.G*, pois esses apresentam a transcrição em coreano em formato de imagem na tela (*hardcoded*), mas com legendas em inglês disponíveis pelo *YouTube*, o que possibilita a extração direta em texto do inglês, mas dificulta alinhamento por entrada.

Para resolver o problema de extração e alinhamento do texto é necessário um software de OCR (*Optical Character Recognition*) com capacidade de reconhecimento de caracteres coreanos (*hangul*) e do alfabeto latino ao mesmo tempo. Assim, buscamos como solução a utilização do software *VideoSubFinder*, desenvolvido por Simeon Kosnitsky em 2007. A

versão utilizada foi a 6.10, com data de atualização em 4 de abril de 2023, e o processamento de imagem para cada vídeo foi realizado.

O *software VideoSubFinder* foi escolhido em grande parte pela sua gratuidade e simplicidade de uso. Após realizado o processamento das imagens, podemos extraí-las em agrupamentos de uma grande imagem populada por todas as entradas encontradas no vídeo. Essas imagens com as legendas agrupadas e prontas para processamento são então carregadas como PDF e podem ser editadas e extraídas como texto por qualquer *software* de edição de PDF, que também realiza um processo de OCR, mas, dessa vez, com imagens muito mais claras. Em nosso caso, realizamos o *upload* dos arquivos PDF ao Google Drive e os transformamos em texto ao abrir o arquivo como um documento editável, utilizando o Google Docs. Dessa forma, obtemos, em texto, todas as entradas de legendas em coreano e as alinhamos às entradas do inglês; depois realizamos a tradução para o português, para obter assim as três línguas alinhadas por entrada de legenda.

Subcorpus 3: entrevistas realizadas com jogadores profissionais (EJ)

Para a realização do *subcorpus* com entrevistas realizadas pelos jogadores profissionais, foram escolhidos conteúdos trazidos pelo canal midiático KORIZON Esports. O canal é conhecido por trazer entrevistas e conteúdo da Coreia do Sul para o público internacional por meio da tradução e legendagem pelo inglês. Co-fundada por Ashley Kang, a página conta desde 2019 com artigos e publicações constantes da liga oficial de *League of Legends* coreana – a LCK – e conteúdo dos campeonatos mundiais.

Outra razão para a escolha do canal é a qualidade e facilidade de acesso às legendas, uma vez que tanto as transcrições do coreano quanto as legendas equivalentes são disponibilizadas diretamente pela plataforma do YouTube. Dessa forma, o alinhamento de entradas é realizado de maneira mais simples.

Os vídeos selecionados para a inclusão inicial no *subcorpus* são os referentes ao Campeonato Mundial de 2023, em entrevistas realizadas com jogadores nativos da Coreia do Sul. Ao total, foram selecionadas oito entrevistas realizadas por repórteres e jogadores coreanos na língua coreana. Uma tradução manual foi realizada para o português com base nos textos da legenda do inglês e do coreano. Os arquivos foram salvos como .txt e divididos por time, jogador, data de publicação e idioma do arquivo.

Compilação e análise do corpus

Para a análise do *corpus*, foi utilizado o *software AntConc*, desenvolvido por Laurence Anthony, em sua versão 4.2.4, publicada em 2023. Para ajudar na visualização das entradas alinhadas, foi utilizado o *software* Excel, da Microsoft, em versão 2310, publicada no ano de 2023. As planilhas do Excel nesse caso assumiram uma função de ajuda para visualização

de dados entre os diferentes idiomas, e a ferramenta da Microsoft não foi utilizada em nenhum momento para a contagem de *tokens*, *types* ou ranqueamento de palavras devido à maior adaptação do *AntConc* para tal tarefa.

Em relação ao ranqueamento de palavras e identificação de termos, a biblioteca *NLTK* – *Natural Language Toolkit* – da linguagem de programação *Python* foi utilizada em sua versão 3.8.1, publicada em 2023, para a filtragem de *stopwords* – palavras de uso comum classificadas como não importantes/sem conteúdo único, como artigos, conectores e similares – do português e do inglês. O processamento dos arquivos foi feito por meio de um *script* em *Python* e com a utilização das bibliotecas “*tokenize.word_tokenize*” e “*corpus.stopwords*” da *NLTK* para a remoção de *stopwords*. Pela falta de suporte à língua coreana dentro da biblioteca *NLTK*, foi escolhido um repositório de *stopwords* da língua coreana criado por Gene Diaz em 2016 e disponível por meio da plataforma *GitHub*⁷.

Resultados

A coleção dos três *subcorpora* resulta em um *corpus* com o total de 297.684 *tokens* e 36.128 *types*. A Tabela 1 detalha o total de *tokens* e *types* por língua.

Tabela 1 – Número de *tokens* e *types* do KPEEC

Língua	<i>Tokens</i>	<i>Types</i>
Coreano	80.379	19.219
Inglês	105.696	8.092
Português	111.609	9.733

Fonte: elaboração própria.

É possível perceber inicialmente uma grande diferença na quantidade de *tokens* e *types* entre a língua portuguesa e a língua coreana. Isso é um padrão recorrente em todos os *subcorpora* e uma possível explicação é a característica aglutinante da língua coreana, que por meio de partículas e indicadores pode restringir ideias diversas em uma unidade lexical. Além disso, a língua coreana pode ser classificada como uma língua *pro-drop*, referente a *pronoun dropping*, que é uma língua em que certas classes de pronomes podem ser omitidas quando podem ser inferidas pragmática ou gramaticalmente (Welo, 2013). Ambas as características citadas podem ser vistas em uma das entradas do *subcorpus* 3 (EJ), conforme o exemplo a seguir.

⁷ Disponível em: <https://github.com/stopwords-iso/stopwords-ko/commits?author=genediazjr>. Acesso em: 10 jan. 2025.

- Texto original em coreano: 팬분들한테 메세지요? [paenbundeulhante mesejiyo?]
 - Texto em inglês: A message to fans?!
 - Texto em português: Uma mensagem para os fãs?!

A mesma entrada conta então com dois tokens no coreano e cinco no português. Primeiramente, no português, a direcionalidade em “para fãs” é representada pela partícula “한테” [hante], que traz a ideia de “para algo/alguém”, e a pluralidade, em “os fãs”, é realizada pelo caractere “들” [deul], que traz a ideia de plural. Além disso, o “요” [yo] no final indica nível de polidez na fala, que gera diferentes *tokens* a depender da terminação formal ou informal utilizada. Devido a essa e outras situações em que o sistema de partículas e caracteres aglutinantes do coreano diminui a quantidade de itens lexicais, o *corpus* da língua coreana é significativamente menor em quantidade de *tokens*.

Após a compilação individual dos três *subcorpora*, criamos uma tabela na ferramenta Excel com abas individuais para cada *corpus* e aba geral com a coleção dos *corpora*, com o intuito de facilitar a identificação comparativa de entradas. Dessa forma, foi possível identificar variações entre as línguas, em uma perspectiva *intracorporis*, e comparação do uso de terminologias do jogo, como itens e campeões, em uma perspectiva *intercorpora*.

Incluímos romanizações (entre colchetes) nas tabelas com caracteres coreanos em suas devidas colunas para facilitar a leitura daqueles que não possuem familiaridade com o *hangul*, sistema de escrita da língua coreana.

Análise intracorporis

O *subcorpus* 1, da documentação de jogos, conta com 25.0974 *tokens* e 31.599 *types* distribuídos entre 694 arquivos, referentes aos 165 campeões e 158 itens existentes no *League of Legends* e 371 cartas existentes no *Legends of Runeterra*.

Tabela 2 – Número de *tokens* e *types* do *subcorpus* 1 (DJ)

Língua	Token	Type
Coreano	69.284	16.214
Inglês	87.190	9.087
Português	94.500	10.176

Fonte: elaboração própria.

O *subcorpus* 1 (DJ) é aquele que apresenta mais *tokens* e *types* dentre todos os *subcorpora* do KPEEC. Ao introduzir os nomes oficiais de campeões, itens e cartas, os seus *types* mais frequentes são relacionados a mecânicas centrais dos jogos, como dano e vida.

Tabela 3 – 10 *types* mais frequentes do subcorpus 1 (DJ)

Inglês		Coreano		Português	
Type	Frequência	Type	Frequência	Type	Frequência
damage	1.740	피해를 [pihaereul]	1579	dano	1.885
seconds	833	수 [su]	767	inimigos	970
enemies	816	마법 [mabeop]	736	s	775
magic	734	입힙니다 [iphipnida]	680	causando	701
enemy	674	초 [cho]	654	mágico	641
dealing	631	기본 [gibon]	467	inimigo	525
attack	468	있습니다 [itseupnida]	460	velocidade	515
speed	437	적에게 [jeokege]	429	vida	493
health	409	시 [si]	404	ataque	468
target	397	적을 [jeokeul]	403	causa	454

Fonte: elaboração própria.

O segundo *type* mais frequente na língua portuguesa, “s”, se dá pelo fato de que uma grande parte do *subcorpus* 1 (DJ) é referente às habilidades dos campeões de *League of Legends*, e parte das explicações das habilidades inclui a quantidade de segundos de duração de uma habilidade. A escolha estilística na documentação de língua portuguesa foi a de abreviar segundos para “s”, enquanto a documentação da língua coreana e da língua inglesa mantiveram as palavras completas: 초 [cho] (quinta posição) e *seconds* (segunda posição).

Assim como os outros *subcorpora*, o número de *tokens* da língua coreana é inferior aos da língua portuguesa, porém o sistema de partículas da língua coreana é mais evidente no *subcorpus* 1 (DJ), uma vez que este segue a forma escrita formal coreana para explicar habilidades e características de elementos do jogo. Na fala, por vezes, tais partículas são omissas.

É interessante também realizar uma pequena divisão dentro do próprio *subcorpus* DJ, entre informações do jogo *Legends of Runeterra* e informações dos campeões e itens do *League of Legends*. Por se tratar de um jogo de cartas que é jogado em turnos, o *Legends of Runeterra* depende de uma maior descrição de ordem dos acontecimentos e capacidades de cada carta. Enquanto isso, o escopo terminológico aparenta ser maior quando se trata de *League of Legends*, um jogo com interações entre jogadores, itens, terrenos e diversos outros fatores em tempo real.

Uma das palavras mais frequentes nesse *subcorpus* é a referente a dano, no sentido de dano a ser causado ou sofrido por personagens. Esse dano pode ser acompanhado de vários adjetivos para descrever o seu tipo e origem (como, por exemplo: mágico, físico, real).

Uma das características de maior relevância dentro dessa temática de dano é a escolha do uso da palavra 피해 [pihae] como equivalente de “dano” no português (“damage” no inglês).

A escolha do substantivo 피해 [pihae] é relevante por diferentes fatores. Em primeiro lugar, pelo fato de muitos jogos possuírem exclusivamente a língua inglesa em sua documentação, existe hoje no léxico coreano a palavra 데미지 [demiji] uma espécie de estrangeirismo existente na forma da adaptação da palavra *damage* do inglês para o sistema coreano – fenômeno esse que chamaremos de “hangulização”. Além disso, há também um termo designado especialmente para o ato de *causar dano*, possivelmente proveniente do verbo do inglês, *deal* (damage): 딜 [dil]. Ambos os termos vindos do inglês aparecem no *subcorpus 2* (CJ). O termo 딜 [dil] também aparece como parte do nome de uma das cinco posições do *League of Legends* em coreano. 윈딜 [wondil], referente à posição de atirador no jogo, é uma contração do termo 윈거리 딜러 [wongeoril dileo]. 윈거리 [wongoeori] significa “longa distância” e 딜러 [dilleo] vem do inglês *dealer*, neste caso, podendo ser entendido como aquele que causa dano à distância.

Apesar de não se de saber exatamente o motivo pelo qual se escolhe utilizar uma palavra exclusiva do léxico coreano, é possível aventar algumas hipóteses: devido a necessidade de maior clareza e acessibilidade no ambiente de documentação que servirá como base do jogo; devido ao desejo de preservar o léxico coreano sem empréstimos ou estrangeirismos do inglês. Outra observação importante sobre o substantivo de dano, 피해 [pihae], é a existência de uma palavra de grafia similar em forma verbal, 피하다 [pihada], com o significado de desviar; esquivar. O verbo também se torna uma homonímia em conjugação no presente declarativo informal: 피해 [pihae].

O *subcorpus 2* (CJ) conta com 26.028 *tokens* e 5.158 *types* distribuídos entre 24 arquivos, referentes aos 8 vídeos divididos entre os três canais citados anteriormente.

Tabela 4 – Número de *tokens* e *types* do subcorpus 2 (CJ)

Língua	Token	Type
Coreano	6.368	2.349
Inglês	10.293	1.647
Português	9.367	1.964

Fonte: elaboração própria.

O *subcorpus* CJ é o segundo maior *subcorpus* em *tokens* e *types* dentre os três *subcorpora* do KPEEC, e nos oferece o uso da língua em contexto de jogo por jogadores profissionais.

Tabela 5 – 10 *types* mais frequentes do *subcorpus* 2 (CJ)

Inglês		Coreano		Português	
Type	Frequência	Type	Frequência	Type	Frequência
flash	122	이거 [igeo]	107	flash	121
let	104	돼 [dwae]	99	vamos	113
ult	84	안 [an]	59	vou	92
good	76	노플 [nopeul]	57	ult	74
go	68	한 [han]	49	podemos	62
look	67	궁 [gung]	46	ei	59
hey	60	수 [su]	44	aqui	56
faker	58	천천히 [cheoncheonhi]	44	bom	55
oner	56	있어 [iteo]	43	pode	52
back	53	나이스 [naiseu]	42	posso	50

Fonte: elaboração própria.

O *subcorpus* CJ propicia a comparação entre diferenças da terminologia utilizada em meios de comunicação de caráter mais oficial e do uso real entre jogadores.

As palavras "faker" e "oner" são os nomes de jogadores da equipe da T1, que é uma das três equipes que foram utilizadas como fonte para coleta de dados. Dessa forma, seria necessária a extração de mais dados de outras equipes e jogadores para uma maior variedade nesse *subcorpus*.

Outra característica a se observar é a frequência do *type* "flash", ocupando o primeiro lugar de frequência tanto no português como no inglês. O *flash* é um feitiço de invocador presente no jogo *League of Legends* que desloca o jogador instantaneamente na direção de seu cursor. Tal feitiço possui grande importância no jogo, pois só pode ser utilizado uma vez a cada 5 minutos e permite realizar diferentes jogadas. Dessa forma, os jogadores profissionais estão constantemente comunicando sobre quem tem esse feitiço disponível ou não. O *type* relacionado ao *flash* em coreano é visto na quarta colocação, atrás apenas de pronomes e bases verbais do léxico coreano, com o *type* 노플 [nopeul]. Iremos aprofundar a análise deste *type* ao realizar uma comparação intercorpora.

O *subcorpus* 3 (EJ) conta com 20.731 *tokens* e 4.497 *types* distribuídos entre 24 arquivos, referentes a oito entrevistas realizadas em 2023 na língua coreana no canal do YouTube da KORIZON *Esports*, conforme citado anteriormente.

Tabela 6 – Número de *tokens* e *types* do *subcorpus* 3 (EJ)

Língua	Token	Type
Coreano	4.759	2.062
Inglês	8.220	1.168
Português	7.752	1.464

Fonte: elaboração própria.

O *subcorpus* 3 (EJ), apesar de ser o menor *subcorpus* dentre os três *subcorpora*, nos permite analisar se há alguma diferença no vocabulário utilizado por jogadores profissionais ao tratar do jogo em um ambiente diferente. Um exemplo simples é o *type* mais frequente do inglês, “*dk*”, referindo-se a “Damwon Kia”, o nome de um dos times do campeonato em que foram realizadas as entrevistas, conforme se verifica na Tabela 7.

Tabela 7 – 10 *types* mais frequentes do *subcorpus* 3 (EJ)

Inglês		Coreano		Português	
Type	Frequência	Type	Frequência	Type	Frequência
dk	59	같아요 [gatayo]	48	sobre	64
worlds	58	그런 [geureon]	47	jogo	64
game	53	많이 [mani]	43	dk	58
think	52	잘 [jal]	36	mundial	56
team	37	수 [su]	35	então	43
kt	35	오늘 [oneul]	29	kt	35
lane	35	저희가 [jeohuiga]	29	final	34
good	34	같습니다 [gatseupnida]	27	bem	33
like	33	같은 [gateun]	25	equipe	31
really	31	할 [hal]	24	contra	30

Fonte: elaboração própria.

Similares às limitações do *subcorpus* 2 (CJ), o *subcorpus* 3 (EJ) também possui um viés perceptível pelos seus *types* mais frequentes, que incluem “*dk*” e “*kt*” entre os 10 *types* mais frequentes do português e inglês. Tanto DK quanto KT são nomes de equipes coreanas que tiveram seus jogadores entrevistados. No total, foram entrevistados jogadores de quatro equipes diferentes.

É possível perceber também nos *types* frequentes do coreano uma maior formalidade na conjugação dos verbos. Verbos no coreano são conjugados e possuem sufixos flexionais para indicar, entre outras características, a “polidez” ou formalidade da conjugação. Dessa forma, dois *types* dentre os 10 mais frequentes demonstram essas conjugações de maior respeito e polidez: *같아요* [*gatayo*] e *같습니다* [*gatseupnida*]. Ambos os *types* são conjugações formais do verbo *같다* [*gatda*], que traz o significado de “parecer” ou “ser igual”. É um verbo muito utilizado em falas mais “oficiais” – tal qual o contexto da entrevista, quando o falante busca evitar realizar afirmações muito concretas. Pode-se comparar à modalização do português nos casos em que o falante usa “acho que” ou “penso que” para realizar afirmações em sua fala. A mesma formalidade e polidez ocorrem no *type* *저희가* [*jeohuiga*], em que *저희* [*jeohui*] é uma versão mais polida e formal de *우리* [*uri*], significando “nós”, seguido pela partícula *가* [*ga*], que marca o sujeito da frase.

As entrevistas foram úteis para visualizar a tradução de termos relacionados ao escopo dos *esports*, mas não limitadas aos termos que são utilizados no decorrer da partida. Um desses exemplos é a terminologia utilizada durante a fase de escolha dos campeões que ocorre previamente às partidas de *League of Legends*. Há a presença de um empréstimo vindo do inglês para o português para se referir a essa fase, por meio da palavra *draft* – referente a esboço; projeto. O ato de selecionar e banir campeões (montar a composição de campeões) antes do início da partida ficou popularmente conhecido em português como o ato de “draftar”, com uma frequência do radical “draft” aparecendo ao total 34 vezes no *subcorpus* em português. Tal *type* não aparece entre os 10 mais frequentes por se tratar do radical (“*draft**”), e não de um *type* único – “*drafts*” e “*draft*” aparecem 12 vezes cada, e outras variações (“*draftamos*”, “*draftando*”, “*draftou*”) compõem as outras 10 ocorrências. Por meio do alinhamento de entradas, foi possível traçar uma relação tradutória do termo com o termo coreano de *밴픽* [*baenpik*], que também é de origem inglesa, de *ban* e *pick*, seguindo a ideia de escolhas e banimentos. Os termos de escolha e banimentos de campeões nessa fase pré-jogo também penetraram a língua portuguesa, com o radical “*pick*” aparecendo 15 vezes no *subcorpus*.

Outro termo presente no *subcorpus* que permeia todas as línguas é a palavra *scrim*, do inglês: significando partidas-treino jogadas com outras equipes como forma de treino ou prática. O termo teve a sua adaptação como substantivo e também passou por um processo de “verbalização” no português, assim como outros empréstimos. O termo também passou

pelo processo de “hangulização” e assumiu a forma de 스크림 [*seukeurim*]. A permeabilidade do termo pode ser vista em entrada a seguir, extraída do *subcorpus* 3 (EJ):

- Texto original em coreano: 이번에 스크림도 몇 번 해보고 하면서 [*ibeone seukeurimdo myeot beon haebogo hamyeonseo*]
 - Texto em inglês: But after scrimming against them a few times,
 - Texto em português: Mas depois de scrimmar com eles algumas vezes,

A Tabela 8 apresenta algumas das equivalências terminológicas encontradas entre as línguas referentes aos elementos discutidos em entrevista que vão além do vocabulário utilizado no decorrer das partidas.

Tabela 8 – Equivalência de termos do *subcorpus* 3

Inglês	Coreano	Português
Champion pool	챔피언 풀 [chaempieon pok]	Champion pool
Scrim	스크림 [seukeurim]	Scrim
Draft	밴픽 [baenpik]	Draft / Picks e bans / Fase de escolhas
Composition / Comp	조합 [johap]	Composição / Comp
Solo queue	솔로 랭크 /솔랭 [solro raengkeu /solraeng]	Solo queue

Fonte: elaboração própria.

Análise intercorpora

Fica evidente algumas mudanças na linguagem utilizada por jogadores em contextos de uso real ao se comparar a terminologia utilizada para se referir a campeões, itens e feitiços do jogo. Um primeiro fenômeno evidente em todas as línguas do *corpus*, e em especial na língua coreana, é a abreviação de palavras.

O campeão Aphelios, do *League of Legends*, apesar de aparecer todas as vezes com o seu nome completo, 아펠리오스 [*apellioseu*] na documentação presente no *subcorpus* 1 (frequência 10) e utilizado uma vez no *subcorpus* 3, não é utilizado em nenhum momento com essa grafia completa no *subcorpus* 2. Em contexto de uso durante as partidas ou transmissões, o nome do campeão foi ressignificado em coreano para 아펠 [*apel*], referindo-se somente ao início da palavra (*Aphel*), a fim de alcançar uma comunicação mais rápida e efetiva. O termo aparece no total 40 vezes no *subcorpus* 2 para a língua coreana, e a grafia completa referente ao nome do campeão não é utilizada uma sequer vez.

O caso de Aphelios é um tipo de abreviação realizada exclusivamente pela língua coreana no escopo encontrado nos *corpus*, uma vez que o nome do campeão é utilizado em sua integridade em todos os *subcorpora* da língua portuguesa. Apesar disso, o fenômeno de abreviação de nomes também existe no português para outros campeões, feitiços e itens. A seguir, é possível visualizar uma tabela com exemplos dessas abreviações presentes no *subcorpus* 2 e sua frequência de uso comparado com a sua forma original. Em casos da ausência de abreviação no português e existência no coreano (como Aphelios), um hífen [-] foi utilizado para demonstrar tal ausência. Vale lembrar que nenhuma dessas abreviações aparece nos *subcorpora* 1 e 3, considerados meios mais oficiais de comunicação.

Tabela 9 – Abreviações no subcorpus 2 (CJ)

Nome em PT e KR	Abreviação em PT e KR	Frequência de uso abreviado KR	Frequência de uso completo KR	Frequência de uso abreviado PT	Frequência de uso completo PT
Aphelios / 아펠리오스	- / 아펠 [ape/]	40	0	-	40
Sejuani / 세주아니	Seju / 세주	12	1	17	2
Renekton / 레넥톤	Renek / 레넥	7	1	7	2
Flash / 점멸*	- / 플	114	2	-	116
Malphite / 말파이트	- / 말프	14	2	-	16
Aatrox / 아트록스	- / 아트	25	0	-	25
Sylas / 사일러스	- / 사일	20	0	-	23
Ultimate / 궁극기	Ult / 궁	92	1	93	-

Fonte: elaboração própria.

Destacado com um asterisco, o termo *flash*, referente a um “feitiço de invocador” dentro do jogo não só é um exemplo de abreviação como também de estrangeirismo ou empréstimo do inglês para a formação de termos nas línguas receptoras. Conforme discutido anteriormente, devido a resignificação ao léxico e eventuais mudanças ortográficas, seria interessante considerar essas inclusões como empréstimos.

Os feitiços de invocador, apesar de terem traduções para cada língua — por exemplo, o feitiço “*smite*”, do inglês, tem a sua tradução para “golpear” no português e “강타” [*gangta*] no coreano — realizam um empréstimo da língua inglesa em seu uso pelos jogadores, como

é o caso de *flash*, que se encontra no *corpus* tanto sendo usado como verbo quanto substantivo:

- Texto original em coreano: 나 플 쓸게 [*na peul sseulge*]
 - Texto em inglês: I'll **Flash** in.
 - Texto em português: Vou **flashar**.
- Texto original em coreano: 플이라도 빼보자 [*peurirado ppaeboja*]
 - Texto em inglês: Let's at least get his **Flash**.
 - Texto em português: Vamos pelo menos gastar o **Flash** dele.

A maior diferença desse estrangeirismo que acontece tanto no coreano quanto no português é a inclusão de uma etapa de abreviação que ocorre no coreano, em que o termo 플 [*peul*] é uma abreviação da grafia completa de 플래시 [*peullaesi*]. Para esse feitiço em específico, é interessante notar que o mesmo é quase somente chamado por seu nome oficial de 점멸 [*jeommyeol*] em arquivos e páginas oficiais do jogo – uma frequência de 7 ocorrências no *subcorpus* 1, e 2 ocorrências no *subcorpus* 2, muito inferior às 114 ocorrências da forma abreviada do inglês.

Além disso, a expressão de negação do inglês, “no”, foi trazida para o coreano também ao indicar a ausência de *flash*, com a colocação mais frequente à esquerda do termo 플 [*peul*] sendo a palavra 노 [*no*] a representação do “não” em inglês; ou seja, 노 플 [*no peul*] alinha-se com “sem flash” no português. Tal colocação ocorre 79 vezes no *subcorpus* 2 e reforça a forte influência do inglês no vocabulário terminológico do coreano no meio dos *esports*.

O feitiço de invocador *teleporte* passa pelo mesmo processo de empréstimo do inglês. Citado pelo seu nome oficial em coreano, 순간이동 [*sunganidong*] 27 vezes no *subcorpus* 1 (DJ), o feitiço é denominado de 텔 [*te*] pelos jogadores, com 32 instâncias de uso no *subcorpus* 2. O termo é uma abreviação do inglês de *teleport*, com a sua grafia inteira no coreano sendo representada como 텔레포트 [*tellepoteu*]. O mesmo empréstimo e abreviação de forma similar também ocorre no português, com o feitiço assumindo a forma de “TP” (21 instâncias no *subcorpus* 2), ou uma verbalização em “tpar” (4 instâncias):

- Texto original em coreano: 아래에 잭스 텔 정면 [*araee jaekseu tel jeongmyeon*]
 - Texto em inglês: Below, Jax **TP'ing** in front.
 - Texto em português: Embaixo, Jax **tpando** na frente

- Texto original em coreano: 나 5 초 뒤 텔 돼, 5 초 뒤 [na 5cho dwi tel dwae, 5 cho dwi]
 - Texto em inglês: I can **Teleport** in 5 seconds, 5.
 - Texto em português: Posso dar **TP** em 5 segundos, 5.

Um exemplo de um empréstimo do inglês que ocorreu para o português, mas não para o coreano é referente à habilidade de jogo *Ultimate*. A abreviação ocorre para ambas as línguas, com o coreano frequentemente utilizando 궁 [gung] para se referir ao termo 궁극기 [gunggeukgi], e o português utilizando “ult” para se referir à habilidade “ultimate”. No *subcorpus* 1, o termo é utilizado em sua grafia completa 77 vezes na língua coreana e 68 vezes no português, porém a sua utilização foi feita de forma abreviada em todas as 186 ocorrências (divididas em 93 para cada língua) no *subcorpus* 2.

Foi possível também encontrar no *corpus* termos – trazendo a “hangulização” do coreano para a escrita romanizada – semelhantes nas três línguas, especialmente os que se referem a elementos internos do jogo na linguagem utilizada pela comunidade. Ou seja, por mais que às vezes haja uma tradução como *flash* para o coreano, o uso real desse termo se mostrou predominantemente similar entre as três línguas. Um outro exemplo prático são as *rotas*, locais do mapa de jogo do *League of Legends* que se dividem em: rota superior, rota do meio e rota inferior. Apesar de a tradução ter sido feita do inglês para o português e o uso da palavra “rota” no *subcorpus* 1 aparecer em 35 ocasiões e em 2 ocasiões no *subcorpus* 3, o termo “lane” – “top lane, mid lane, bottom/bot lane” – é mais abertamente aceito e utilizado pela comunidade, conforme inclusão do termo *lane* na língua portuguesa em 37 vezes, e a os termos *top*, *mid* e *bot* somando juntos 110 ocasiões divididas entre os *subcorpora* 2 e 3.

A Tabela 10 mostra alguns termos que, apesar de não necessariamente serem a tradução oficial adotada pelo jogo, seguem o mesmo padrão de similaridade de uso (dado por maior frequência e equivalência no *corpus*) entre os três idiomas:

Tabela 10 – Termos similares entre as três línguas

Coreano	Inglês	Português
미드 [mideu]	Mid	Mid
탑 [tap]	Top	Top
바텀 [bateom]	Bot	Bot
플래시 [peullaesi]	Flash	Flash
다이브 [daibeu]	Dive	Dive
정글 [jeonggeul]	Jungle	Jungle
바론 [baron]	Baron	Baron

Fonte: elaboração própria.

Considerações finais

O objetivo deste trabalho foi explorar um *corpus* paralelo que introduzisse a linguagem específica de jogos do coreano à comunidade brasileira em seu contexto de uso real e servisse como referência em uma área de conteúdo escasso. Dessa forma, o objetivo foi parcialmente atingido, uma vez que demonstra a diferença entre a linguagem oficial e a utilizada em contextos reais, e estabelece um pequeno corpus que pode ser expandido futuramente.

Apesar da criação desse *corpus* demonstrar a possibilidade de análise linguística entre as três línguas, ele é, em sua natureza, um *corpus* com algumas limitações. A maior delas, conforme vista no processo de coleção e compilação de textos, é a falta de bidirecionalidade no corpus, em que há pouco conteúdo de língua portuguesa do meio traduzido para a língua coreana. Além disso, a própria falta de ressignificação dos conteúdos coreanos diretamente para a língua portuguesa e a necessidade da inclusão da língua inglesa como idioma intermediário demonstra a necessidade de expansão do domínio para possibilitar o estudo mais aprofundado de interação entre as línguas.

O presente *corpus* também apresenta restrição do ponto de vista de variedade de recursos, de tal forma que apenas 3 times, 1 canal de notícias e 1 empresa de jogos foram contemplados para a sua criação.

Apesar disso, o presente trabalho apontou diferenças entre a linguagem considerada oficial, ou seja, provinda dos manuais e arquivos dos jogos e a de uso real pela comunidade de jogadores. Em ocasiões futuras, é desejável a criação de um *corpus* ainda mais diverso para possibilitar a comparação entre línguas dentro do domínio dos *esports*. A inclusão de traduções para o coreano de conteúdos em português seria um excelente começo para impulsionar essa cooperação linguística intercultural e possibilitar maior entendimento por parte da comunidade brasileira no universo dos *esports* coreanos e vice-versa.

Espera-se que o *corpus* sirva de referência para pesquisas futuras na área, mesmo que não usado em sua forma íntegra, mas que assuma a posição de agente motivador para demonstrar que há interesse no desenvolvimento de uma maior cooperação dentro do domínio dos *esports* e de que é possível tornar mais acessível um nicho terminológico que muitas vezes foge dos livros tradicionais de ensino de línguas. A sua aplicabilidade prática pode ser tanto em nível profissional, para os profissionais da área utilizarem como possível glossário ou referência ao incluir jogadores sul-coreanos em times falantes do português ou inglês, quanto para o uso de criadores de conteúdo, possibilitando uma legendagem mais rápida, prática, e até mesmo possivelmente automatizada – por meio de treinamento de tradutores automáticos com o corpus como *dataset*.

Referências

ALMEIDA, G. M. B.; CORREIA, M. **Terminologia e corpus: relações, métodos e recursos**. In: Stella E. O. Tagnin; Oto Araújo Vale. (Org.). *Avanços da Linguística de Corpus no Brasil*. 1 ed. São Paulo: Humanitas/FFLCH/USP, 2008, v. 1, p. 63-93.

ANTCONC. 4.2.4. [Tokyo]: Laurence Anthony, 2001. Software. Disponível em: <https://www.laurenceanthony.net/software/antconc/>. Acesso em: 13 nov. 2023.

BAKER, M. *Corpus Linguistics and Translation Studies — Implications and Applications*. **Text And Technology**, [S.L.], p. 233, 1993. John Benjamins Publishing Company. <http://dx.doi.org/10.1075/z.64.15bak>.

DAYRELL, C. O USO DE CORPORA PARA O ESTUDO DA TRADUÇÃO: objetivos e pressupostos. **Tradução em Revista**, [S.L.], v. 2015, Especial, p. 87-102, 29 out. 2015. Faculdades Católicas. <http://dx.doi.org/10.17771/pucrio.tradrev.25346>.

DIAZ, G. **Stopwords Korean (KO)**. 2016. Disponível em: <https://github.com/stopwords-iso/stopwords-ko>. Acesso em: 13 nov. 2023.

DPLUS KIA. **Canal do YouTube**. Disponível em: https://www.youtube.com/@DplusKIA_LOL. Acesso em: 12 nov. 2023.

INSTITUTO NACIONAL SUL-COREANO DE LÍNGUA COREANA. **Translation Guidelines for Building a Korean-Foreign Language Parallel Corpus**. [S. l]: Hawoo Publishing, 2023. 348 p. ISBN 979-11-6748-090-3 93710.

JOHANSSON, S. *Seeing Through Multilingual Corpora: On the Use of Corpora in Contrastive Studies*. John Benjamins Publishing Company, 2007

FINATTO, M. J. B.; KRIEGER, M. G. **Introdução à Terminologia**: teoria e prática. São Paulo: Contexto, 2004.

NLTK. 3.8.1. [S. l.]: NLTK Project, 2005. Software. Disponível em: <https://www.nltk.org>. Acesso em: 13 nov. 2023.

OLOHAN, M. **Introducing Corpora in Translation Studies**. Londres: Routledge, 2004. 232 p. ISBN: 9780203640005.

T1. **Canal do YouTube**. Disponível em: <https://www.youtube.com/@SKTT1>. Acesso em: 12 nov. 2023.

VIDEOSUBFINDER. 6.10. [S. l.]: Simeon Kosnitsky, 2007. Software. Disponível em: <https://sourceforge.net/projects/videosubfinder/>. Acesso em: 5 nov. 2023.

WELO, E. **Encyclopedia of Ancient Greek Language and Linguistics**. [S. l.]: Brill, 2013. General Editor: Georgios K. Giannakis. Disponível em: http://dx.doi.org/10.1163/2214-448X_eagll_COM_00000254. Acesso em: 13 nov. 2023.

Sobre os autores

Vítor Amaral Córdoba

Orcid: <https://orcid.org/0009-0008-5724-2189>

Bacharel em Línguas Estrangeiras Aplicadas ao Multilinguismo e à Sociedade da Informação (LEA-MSI) pela Universidade de Brasília (UnB).

Thiago Blanch Pires

Orcid: <https://orcid.org/0000-0002-0060-6075>

Professor adjunto do Departamento de Línguas Estrangeiras e Tradução (LET) e do Programa de Pós-Graduação em Linguística (PPGL) vinculados ao Instituto de Letras (IL) da Universidade de Brasília (UnB).

Recebido em fev. 2025

Aprovado em mai. 2025