A "OTIMIZAÇÃO DOS MECANISMOS DE BUSCA" (SEO) COMO FERRAMENTA DE COLETA AUTOMATIZADA DE DOCUMENTOS PARA ELABORAÇÃO DE CORPORA*

Jean-Claude Miroir[1]

Resumo

Este trabalho fundamentou-se no quadro teórico desenvolvido nos "Estudos da *Tradução embasada em Corpora*" (*Corpus-Based Translation Studies*) (Baker, 1993; Laviosa, 2002; Olohan, 2004; Zanettin, 2012; Vania e Tagnin, 2015). A metodologia de "otimização dos mecanismos de busca" (SEO – *Search Engine Optimization*) (Enge *et al.*, 2012) foi desenvolvida para elaborar, sem conhecimentos especializados em informática, corpora comparáveis (Olohan, 2004; Zanettin, 2012) de forma rápida e confiável. Com efeito, este estudo visou reduzir de forma significativa o "ruído" (conjunto de documentos irrelevantes) gerado na coleta automatizada de documentos por meio de ferramentas específicas. Os procedimentos da "otimização dos mecanismos de busca" (SEO) implementados permitiram o aproveitamento de todo o potencial dos recursos disponíveis nas configurações da "pesquisa avançada" dos mecanismos de busca como o *Google Search Engine* e dos "operadores de pesquisa". A melhora da qualidade dos resultados gerados por essa metodologia de pesquisa agilizou a produção dos corpora customizados (Perrotti-Garcia, 2005) ou dos DIY [*Do-It-Yourself*] Corpora (Zanettin, 2002), criados a partir da conversão em lotes dos corpora comparáveis. Esses procedimentos rápidos e eficientes resultaram na (re)valorização do uso desses tipos de corpora *"ad hoc"* no fluxo de trabalho do tradutor profissional e do aprendiz, contribuindo no melhoramento da qualidade do produto final: a tradução.

Palavras-chave: *Tradução embasada em Corpora*. Otimização dos mecanismos de busca (SEO). Coleta automatizada de documentos. Corpus comparável. Corpus customizado.

Search Engine Optimisation (SEO) as a tool for the automated collection of documents for developing the corpora

Abstract

This work was based on the theoretical framework developed in the Corpus-Based Translation Studies (Baker, 1993; Laviosa, 2002; Olohan, 2004; Zanettin, 2012; Vania e Tagnin, 2015). The methodology of "search engine optimisation" (SEO) (Enge et al. 2012) was developed to prepare comparable corpora without expertise in computer science (Olohan, 2004; Zanettin, 2012) quickly and reliably. Indeed, this study aimed to significantly reduce the "noise" (set of irrelevant documents) generated in the automated collection of documents by means of specific tools. The "search engine optimisation" (SEO) procedures implemented allowed the use of the full potential of available resources in the search engines "advanced search" settings such as Google Search Engine and "search operators". Improving the quality of the results generated by this research methodology facilitated the production of customised corpora (Perrotti-Garcia, 2005) or DIY (Do-It-Yourself) Corpora (Zanettin, 2002), created from the conversion in batches of comparable corpora. These fast and efficient procedures resulted in the (re)valuation of the use of these types of "ad hoc" corpora in the workflow of the professional translator and the apprentice, contributing to the improvement in quality of the final product: the translation.

Keywords: Corpus-Based Translation. Search Engine Optimisation (SEO). Automated collection of documents. Comparable corpus. Customised corpus.

* Gostaríamos de agradecer à Capes (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) pelo financiamento dos anais da VII Escola Brasileira de Linguística Computacional e do XIII Encontro de Linguística de Corpus, processo nº 3472/2015-87.

[1] Doutor em Literatura Brasileira pela Universidade de Brasília (2013). Professor Adjunto do curso de Letras-Tradução-Francês do Departamento de Línguas Estrangeiras e Tradução (LET) da Universidade de Brasília (UnB). CV Lattes: http://lattes.cnpq.br/3299590986801782 - Contato: jc.unb@hotmail.com

1. Introdução

A Linguística de Corpus tornou-se uma aliada valiosa para o tradutor. Além de seu papel imprescindível para o estudo e ensino das línguas naturais (SARDINHA, 2000), ela oferece também um acesso mais amplo e seguro à termologia de especialidade relacionada a determinada área de conhecimento à qual o texto a ser traduzido pertence. Todavia, neste artigo objetiva-se analisar um aspecto mais específico e prático dos "Estudos da tradução", isto é, o processo tradutório propriamente dito que, conforme Sardinha (2002, p. 15), "têm muito a ganhar com um contato maior com a Linguística de Corpus". Esse contato manifesta-se concretamente nos "Estudos da *Tradução embasada em Corpora*" (*Corpus-Based Translation Studies*) (Baker, 1993; Laviosa, 2002; Olohan, 2004).

No entanto, vale ressaltar que essa abordagem teórica foi influenciada pela própria Linguística de Corpus e pelos "Estudos Descritivos da Tradução" (*Descriptive Translation Studies* – DTS). Seu principal foco permanece sendo os estudos de traduções, para o inglês, de textos, essencialmente, oriundos dos espaços literários globais. As ferramentas computacionais da Linguística de Corpus são empregadas para pesquisar o "produto" que representa o "texto traduzido". Esse produto adquiriu suas *letras de nobreza* graças a pesquisadora Mona Baker (1993), pois o "texto traduzido" ficou afastado, muito tempo, da composição de corpora, em razão de sua falta de representatividade da (e na) língua de chegada (LC), bem como, os textos produzidos por aprendizes de uma língua estrangeira. A análise descritiva visa determinar o processo de tradução subjacente de inúmeros tradutores para identificar e definir padrões estilísticos, independentemente do par de línguas em contato, característicos do texto que possui doravante o estatuto acadêmico de "texto traduzido". Além de seu interesse para pesquisa universitária, o uso de corpora em aulas de línguas estrangeiras foi, de um lado, uma das primeiras aplicações didáticas da Linguística de Corpus. De outro lado, autores como Zanettin, (2012) e Vania e Tagnin (2015) refletiram sobre seu papel na formação de tradutores, no ensino da tradução.

A justificativa deste trabalho, por sua vez, é oferecer aos tradutores profissionais e aprendizes maior e melhor acesso à *Tradução embasada em Corpora* (TEsC) e suprir determinadas limitações dos recursos existentes. Com efeito, a maioria dos grandes corpora de referências (*Ready-made corpora*) está disponível em língua inglesa. Em contexto tradutório com pares de línguas diferentes, por exemplo, neste estudo, o português do Brasil (pt-BR) e o francês (fr-FR), eles podem ser úteis utilizando o inglês apenas como idioma pivô. Os corpora acessíveis *online* nesses dois idiomas (pt-BR e fr-FR), geralmente de menor extensão e equipados de interfaces de consultas com recursos limitados, não proporcionam sempre resultados satisfatórios para ajudar o tradutor nas suas tomadas de decisão.

Portanto, o argumento consiste em apresentar um método coerente para incentivar os tradutores a criarem brevemente seus próprios corpora *ad hoc* ou DIY[2] *corpora* (Zanettin, 2002; 2012) para suprir essa carência. O foco desse estudo centra-se no aspecto específico da coleta automatizada por meio de uma ferramenta de análise para a otimização de mecanismos de busca o *SEOquake* (www.*SEOquake*.com), uma extensão (*plug-in*) disponível para os navegadores *Firefox, Chrome* e *Opera*.

[2] Do-It-Yourself

No âmbito deste artigo, a metodologia de TEsC exposta focaliza-se na elaboração de dois pares de corpora (comparáveis e customizados), um na língua de partida (LP) e outro na língua de chegada (LC). Viana e Tagnin (2011, p. 328) definem o corpus comparável como um "corpus composto por textos originais numa língua e traduções nessa mesma língua. Tem por objetivo comparar a linguagem produzida por falantes nativos ou fluentes e por tradutores." Vale ressaltar que, neste estudo, as "traduções nessa mesma língua" não serão consideradas, pois devem fazer parte de outro tipo de corpus para outras finalidades de pesquisa ou compor um corpus paralelo. O corpus customizado é definido por Perrotti-Garcia (2005, p. 63), como um "corpus montado pelo próprio usuário, a partir da coleta de textos selecionados por este". Os textos do corpus customizado, ao contrário dos do corpus comparável, são preparados e convertidos no formato "texto simples" (*plain text*), para serem explorados com um concordanciador, como o *AntConc*.

Devido a questões de espaço deste artigo, as respectivas explorações desses dois tipos de corpus, nas línguas de partida e da chegada, durante o processo de TEsC, propriamente dito, serão expostas de forma resumida, no final do trabalho (na Fase C).

2. Revisão da literatura

A Linguística de corpus, uma das áreas de pesquisa em linguagem mais ativas e produtivas nos últimos anos, influenciou significativamente o ensino das línguas estrangeiras e maternas. Assim, segundo Sardinha (2000, p. 358), os trabalhos acadêmicos em Linguística de Corpus são essencialmente empíricos, pois analisam os padrões reais de uso das línguas em textos naturais e autênticos, disponíveis (ou convertidos) em formato eletrônico (HTML, PDF, etc.). Em consequência, o autor ressalva que o uso extensivo de ferramentas computacionais nas análises textuais permite automatizar processos iterativos e pesquisar amplas coletâneas de textos (corpora) por meio de técnicas, não apenas quantitativas, mas também, qualitativas.

A aplicação dos resultados de pesquisas universitárias em Linguística de Corpus na sala de aula subordina-se ainda a, pelo menos, quatro fatores: (1) a formação dos professores; (2) o uso de dicionários; (3) a Internet; (4) os equipamentos computacionais. Supondo que os aspectos técnicos (3) e (4) sejam satisfatórios, a maior dificuldade concentra-se no ponto (1), ou seja, a falta de treinamento para explorar os corpora em sala de aula como complementos eficientes e produtivos dos tradicionais dicionários mono ou bilíngues (em papel ou eletrônicos) (2). Aliás, conforme Sardinha (2000, p. 329), a associação do meio acadêmico especializado em Linguística de Corpus com editoras – por exemplo, a Universidade de Birmingham e a editora *Collins* – produziu inúmeros dicionários, gramáticas e livros didáticos para o ensino de idiomas (basicamente inglês), embasada em corpus, como o COBUILD (*Collins Birmingham University International Language Database*).

Assim, dado que o material didático provém de corpora, Cimermanová (2012) indagou sobre as possibilidades de se substituir em sala de aula os dicionários por corpora e concluiu que:

Ambos, dicionários e corpora são fontes valiosas para o ensino de línguas e combiná-los em sala de aula pode ajudar os alunos a desenvolver suas habilidades de comunicação, para construir sua autoconfiança, sua autoestima linguística, ensiná-los a assumir riscos e isso pode ter uma influência positiva sobre a sua competência linguística (e social).[3] (CIMERMANOVÁ, 2012, p. 73, tradução minha)

[3] "Both, dictionaries and corpora are valuable source for language teaching and combining them in the classroom can help learners to develop their communication skills, to build their self-confidence, builds their language ego, teaches them to take risks and this can have a positive influence on their language (and social) competence." (CIMERMANOVÁ, 2012, p. 73)

A *Tradução embasada em Corpora* (TEsC) não teria existido sem os instigantes desenvolvimentos precursores da Linguística de Corpus aplicados ao ensino de línguas. Percebe-se, em conformidade com os autores citados acima, que o tradutor encontra-se em situação similar à: (1) do professor de línguas sem formação específica em Linguística de Corpus; (2) do aluno quando pode receber uma influência positiva da exploração de corpora na sua prática tradutória quotidiana. Um dos enfoques positivos destacável, neste trabalho, condiz com a capacidade dos corpora suprir as lacunas (não todas) dos dicionários, glossários e/ou bancos de dados terminológicos em situação de tradução especializada, por exemplo, propondo soluções de forma rápida e segura.

Os estudos atualmente disponíveis se referem, com maior frequência, ao ensino da língua inglesa com os maiores corpora disponíveis *online*, conforme a experiência descrita, anteriormente, por Cimermanová, na República Eslovaca. Alguns exemplos de "grandes corpora" são o "*British National Corpus*" (BYU-BNC – versão gratuita do BNC) (http://corpus.byu.edu/bnc/) ou o "*Corpus of Contemporary American English*" (COCA) (http://corpus.byu.edu/coca) que possibilitam a realização de buscas avançadas e aprimoradas. A exploração desses corpora tornou-se acessível para o professor de língua, o aluno e o tradutor, pois possuem interfaces ergonômicas (aliás, as mesmas) que deixaram de ser reservadas aos pesquisadores em Linguística de Corpus. Para o português, o "Corpus Brasileiro" (LAEL-PUC-SP) foi transferido para a plataforma *Sketch Engine* (acesso pago) (www.sketchengine.co.uk/corpus-brasileiro).

A elaboração de corpora comparáveis *ad hoc* ou DIY [*Do-It-Yourself*] corpora (Zanettin, 2002; 2012) na língua de partida (LP) e de chegada (LC) impõe-se como solução alternativa e eficaz para suprir essa carência, especificamente, para projetos de tradução que não envolvem a língua inglesa. Porém, essa observação não significa de projetos que contam com o inglês como LP ou de LC são descartados dessa metodologia. Zanettin apresenta o DIY, da seguinte forma:

Um DIY web corpus pode ser caracterizado da seguinte forma: (1) É uma coleção de documentos da Internet, ou mais precisamente de páginas da Web em HTML; (2) É criada *ad hoc*, como uma resposta a um texto específico a ser traduzido; (3) É um corpus aberto. Mais material pode ser adicionado se for necessário; (4) É descartável [...] ou virtual [...]; (5) Não é destinado a fazer parte de um corpus mais permanente, e pode ser eliminado logo que a tradução estiver concluída. Permissões de direitos autorais não são necessárias; (6) Como no caso dos "textos paralelos" pode ser um corpus comparável bilíngue ou monolíngue na língua de chegada.[4] (ZANETTIN, 2002, p. 4 do arquivo PDF, tradução minha).

[4] A DIY web corpus can be characterized as follows: - it is a collection of Internet documents, or more precisely of web pages in HTML. - it is created ad hoc as a response to a specific text to be translated. - it is an open corpus. More material can be added as the need arises. - it is disposable [...] or virtual [...]. It is not destined to be part of a more permanent corpus, and can be disposed of as soon as the translation is completed. Copyright permissions are not required. - like "parallel texts" it can be either bilingual comparable or target monolingual.(ZANETTIN, 2002, p. 4)

No entanto, a coleta de textos pertinentes representa sempre uma etapa sensível, pois proporciona, em geral, certo desconforto no tradutor, em razão do caráter fastidioso e demorado desse processo manual. Em contrapartida, pode ser considerado mais seguro do ponto de vista da relevância do material selecionado em relação aos critérios do texto de partida (TP) a ser traduzido. O uso de robôs de coleta automatizada de textos ou páginas da Web (ou partes delas), por exemplo, *BootCat Toolkit* (Bernardini, 2004), agilizam sobremaneira o processo. No entanto, de acordo com as configurações da ferramenta, eles podem coletar, por exemplo, informações irrelevantes gerando "ruídos" que, em função de sua importância, obrigam o tradutor a reduzi-los manualmente ou de forma semiautomática. A aplicação de técnicas de "limpeza" se faz necessária para não prejudicar a pertinência dos corpora criados de forma rápida, embora não sendo fatalmente descartáveis. Ao contrário da afirmação de Zanettin, já mencionada, esses tipos de corpora podem constituir um corpus de referência (ou parte dele), dentro de uma área de conhecimento, na qual o tradutor atua.

Qual seria, no momento, a solução rápida e segura para coletar textos em adequação com as características temáticas e terminológicas do TP?

A fim de responder a essa pergunta, o presente estudo apresenta um procedimento de coleta automatizada embasado na "Otimização dos mecanismos de busca" (*Search Engine Optimization* – SEO). Essa otimização é definida como um "processo de elevação do número de visitantes de um site atingindo uma classificação alta nos resultados de busca retornados por um mecanismo de busca." (Enge *et al.,* 2012, p. 743). Essa ferramenta foi inicialmente projetada para os desenvolvedores da Web, para os profissionais de marketing, entre outros, e doravante explorada por tradutores, neste ensaio. Os mecanismos de busca (*Google Search Engine, Microsoft Bing Search, Yahoo!, DuckDuckGo, etc*.) tornaram-se interfaces imprescindíveis no processo tradutório. Os algoritmos de classificação dos resultados gerados pelo SEO garantem certa aderência em relação à "expressão de busca" (EB), se for necessário, devidamente configurada por meio dos "operadores de pesquisa" (*site:, link:, *, " ", +,* -, OR, AND, etc.) e das funcionalidades avançadas próprias a cada mecanismo de busca (língua, país, tipo de arquivo, data, etc.). As mesmas "páginas de resultados de uma busca" (*Search Engine Results Page* – SERP) são disponíveis para uma coleta manual ou assistida com uma ferramenta de SEO. A diferença reside no modo de transferir os arquivos relevantes para a unidade de armazenamento, como o disco rígido (HD) do computador do tradutor e/ou discos virtuais nas "nuvens". No primeiro caso, o manual, o tradutor vai baixar manualmente cada documento de forma unitária. Quanto ao segundo caso, o automático, os links de acesso aos arquivos são transferidos em lotes (*batch*), em um arquivo do tipo .CSV (*comma separated values*), utilizados para trocar dados com um banco de dados ou uma planilha eletrônica (*Microsoft Excel* ou *Libre Office Calc*).

A missão inicial dos mecanismos de busca é a de oferecer rapidamente para seus usuários os resultados mais relevantes e atualizados possíveis, de acordo com a expressão de busca digitada e/ou apropriadamente configurada (ver o item seguinte de metodologia). Por conseguinte, segundo Enge *et al.* (2012, p. 18), "os mecanismos de busca investem uma grande quantidade de tempo, energia e capital para melhorar sua relevância". Essa preocupação tem uma grande vantagem para o usuário "comum" e, mais ainda, para o tradutor que pretende elaborar de forma autônoma e segura seus próprios corpora em função de suas necessidades e das exigências de qualidade de seus clientes e de prazos exíguos:

Como o sucesso dos mecanismos de busca depende tanto da relevância de seus resultados de pesquisa, as manipulações da classificação dos mecanismos de busca que geram resultados irrelevantes (geralmente chamados de *spams*) são tratadas com bastante seriedade. Cada grande mecanismo de busca emprega uma equipe de pessoas que se concentra exclusivamente em detectar e eliminar *spams* de seus resultados de pesquisa. (ENGE *et al.*, 2012, p. 18)

Os sites que manipulam seus dados (método *Black Hat*) para chegar ao topo das páginas de resultados (SERP) são penalizados pelos mecanismos de busca, sendo desclassificados ou retirados do índex (ou seja, "invisíveis") durante um determinado prazo (por exemplo, o caso da firma alemã, BMW, em 2006). A classificação dos resultados de busca garante, por meio dessas precauções de segurança, certa aderência em relação à expressão de busca. Porém, de acordo com Ferneda (2012, p. 18), a "principal dificuldade dos usuários está em predizer, por meio de sua expressão de busca, quais os termos que foram usados para representar os documentos que satisfarão sua necessidade".

Assim para reduzir os riscos de ruídos inerentes a essa "predição" relativa à escolha e à combinação dos termos de busca ou "palavras-chave" (*Key words*), a metodologia desenvolvida neste estudo embasa-se na extração terminológica dos termos (*N-Grams*) significativos do texto de partida (TP). Todavia, vale ressaltar que o "fato de um termo utilizado na expressão de busca aparecer na representação de um documento não significa que este documento seja relevante para a necessidade do usuário" (FERNEDA, 2012, p. 18).

No item seguinte será apresentada a metodologia para implementar na prática tradutória do tradutor profissional ou aprendiz o processo de *Tradução embasada em Corpora* (TesC), com o apoio da tecnologia de SEO.

3. Metodologia

A metodologia de *Tradução embasada em Corpora* (TEsC), com o apoio da tecnologia de SEO, apresentada neste trabalho, fundamenta-se na proposta de versão – do português do Brasil (pt-BR) para o francês (fr-FR) – de um artigo científico de enologia, "Vinhos espumantes: métodos de elaboração" (CALIARI *et al.*, 2013). Esse texto de partida (TP) é composto, após limpeza, de 1109 *types* (nº de palavras únicas) e de 4511 *tokens* (nº de ocorrências de todas as palavras). A elaboração dos corpora (comparáveis e customizados) é constituída de duas fases distintas. A "Fase A" corresponde a sequência de etapas para a elaboração dos corpora na língua de partida (LP: pt-BR) e a "Fase B" diz respeito elaboração dos corpora na língua de chegada (LC: fr-FR).

A "Fase A" e a "Fase B" são bastante similares, cada uma composta em cinco etapas, ou seja, um total de dez etapas. São elas:

Quadro 1 – As duas fases de elaboração dos corpora na LP e na LC
Fonte: elaborado pelo autor

Fase A – Elaboração dos corpora na LP: pt-BR		Fase B – Elaboração dos corpora na LC: fr-FR	
Etapa	Descrição	Etapa	Descrição
1A	Extrair as "palavras-chave" do TP	1B	Elaborar uma lista de "palavras-chave" na LC com base das "palavras-chave" do TP: "hipóteses de tradução"

2A	Coletar os textos pertinentes a partir dessas “palavras-chave” por meio da SEO (arquivo .csv)	2B	Coletar os textos pertinentes a partir dessas “palavras-chave” por meio da SEO (arquivo .csv)
3A	Baixar em lote (*batch*) os arquivos selecionados (links da SEO)	3B	Baixar em lote (*batch*) os arquivos selecionados (links da SEO)
4A	Organizar os corpora comparáveis na LP	4B	Organizar os corpora comparáveis na LC
5A	Preparar os corpora customizados na LP	5B	Preparar os corpora customizados na LC

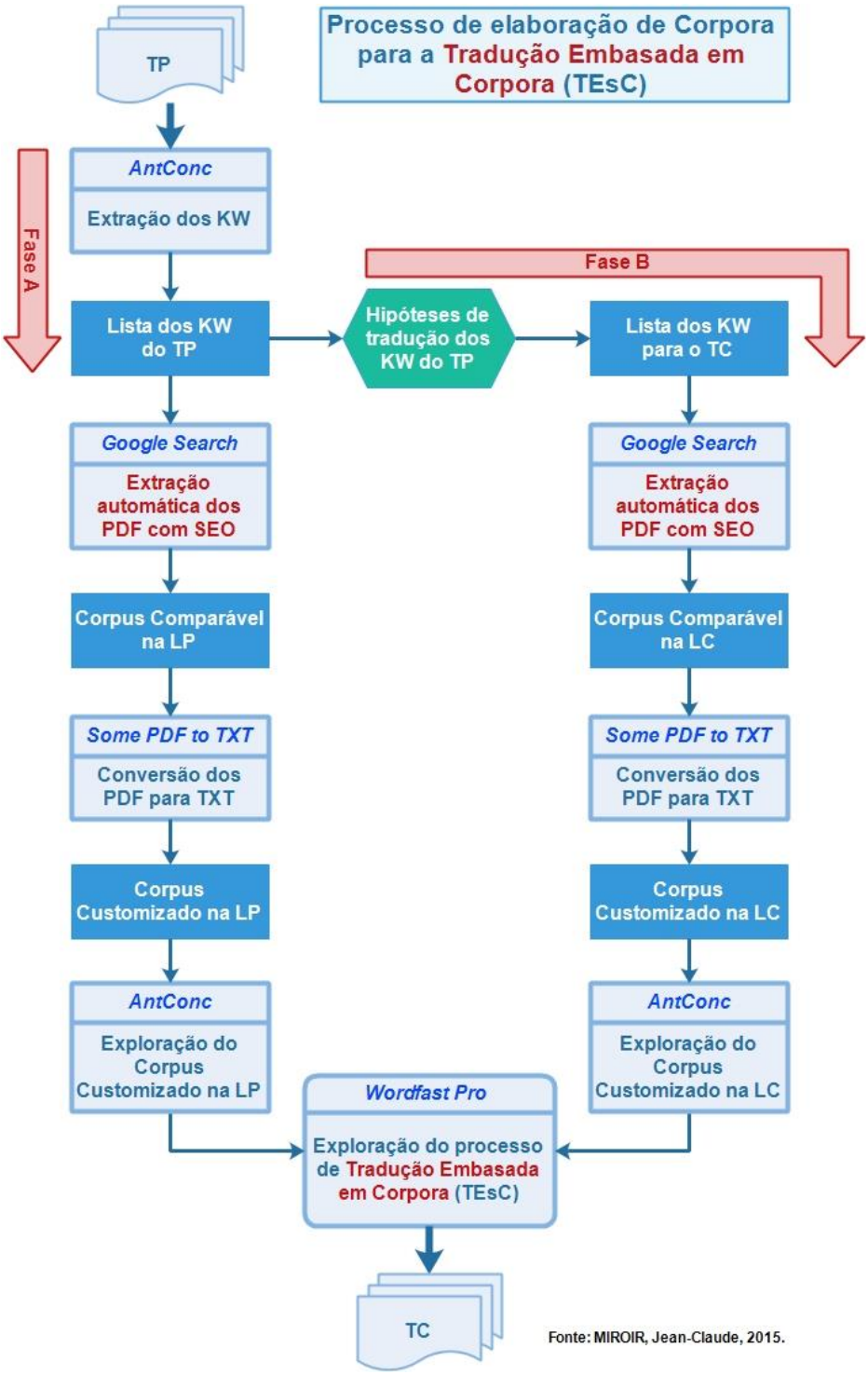


Figura 1 – Processo de elaboração de Corpora para a TEsC

Fonte: elaborado pelo autor

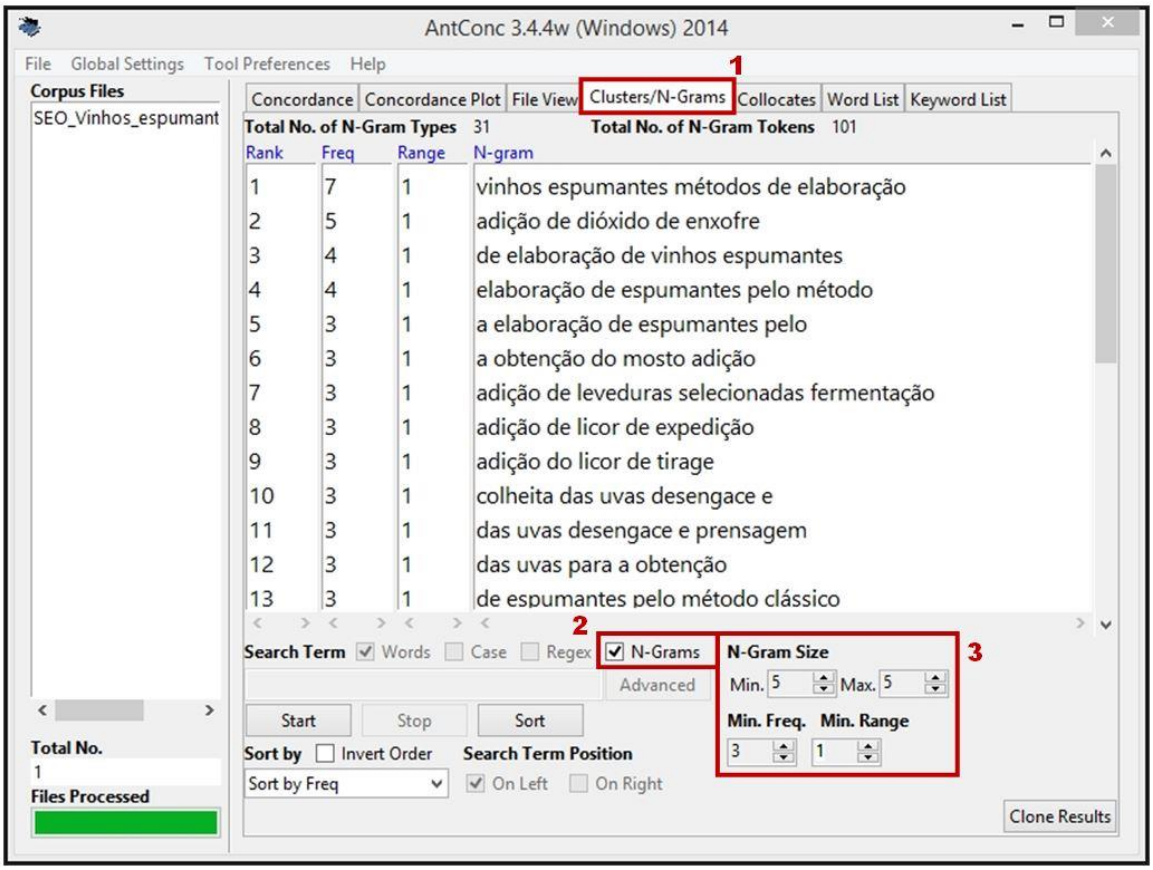
Fase A – Elaboração dos corpora na LP

Etapa 1A. Extrair as “palavras-chave” do TP

Essa etapa é importante para definir a relevância dos textos que serão selecionados automaticamente, pois a coleta é orientada pelas "palavras-chave" (*KeyWords* – KW) extraídas diretamente do TP com o concordanciador *AntConc*. Os concordanciadores são sistemas computacionais de leitura que só aceitam os arquivos de "texto simples" (*plain text*), na extensão .TXT, ou seja, com apenas a base textual sem formatação nem imagens. O TP usado para esse estudo foi convertido com o programa *Some PDF to TXT Converter* (www.somepdf.com). O arquivo criado, com a extensão .TXT, passou por uma limpeza básica para remover, com a função "Substituir" do Bloco de notas: (1) nomes dos autores; (2) credenciais dos autores em notas de rodapé; (3) trechos redigidos em língua diferente da língua principal do texto (*abstract*, citações, etc.); (4) cabeçalhos e rodapés; (5) referências bibliográficas.

Após a limpeza e a importação desse arquivo no *AntConc*, a ativação da ferramenta "*Word List*" permitiu determinar as características do TP (1109 *types* e 4511 *tokens*). O tradutor ativa, em seguida, a ferramenta *N-Grams*, passos [1] e [2], na Figura 2, que extrai do texto analisado os conjuntos de "*N-Grams*", isto é, compostos de N-palavras [ver tabela], que aparecem com uma frequência mínima preestabelecida (*Min. Freq*.) [3] dentro de um número mínimo de textos desse corpus (*Range*) [3]. Assim, no caso da extração terminológica de um único TP, o "Range", nesta etapa, é configurado com o valor "1".

Figura 2 – Processo de extração dos KW (5 *N-Grams*) com o *AntConc*



Fonte: elaborado pelo autor

Foram definidas três extrações sucessivas: A, B e C. A extração A é de maior extensão (*5-Grams*), a B de extensão intermediária (*4-Grams*) e a C de menor extensão (*3-Grams*), que serão exploradas na etapa seguinte par coletar os textos pertinentes do corpus. O valor "*min*." e "*max*." dos *N-Grams*, [3] na Figura 2, são idênticos para que os arquivos de resultados do *AntConc* tornem-se mais fáceis de interpretar e analisar. A regulagem do parâmetro de frequência mínima (*Min. Freq*.) [3] adéqua-se à extensão do TP (nº de *tokens* total) e ao número de ocorrências de palavras-chave (KW) extraídas pelo concordanciador *AntConc* do TP. Assim, quando "*Min. Freq.*" aumenta a filtragem se torna mais restritiva; logo, o número de palavras-chave resultante diminui. O tradutor regula esse valor para receber um número de palavras-chave significativo para realizar sua coleta. Definem-se os "*N-Grams*" como conjunto de N palavras que se associam com determinada frequência dentro de um texto ou corpus.

Os resultados A, B e C são salvos em arquivos de textos, com a extensão .TXT, automaticamente gerados pelo concordanciador *AntConc* : [*File > Save Output to Text File> nome do arquivo.txt*]. A nomeação dos arquivos pode ser a mais explícita possível para o tradutor, por exemplo, os resultados A (5-*Grams*):

AntConc _BR_NGrams_5x5_F3_R1_2015-10-16.txt
Em que:

AntConc:	Resultados oriundos da ferramenta *AntConc;*
BR:	língua do conteúdo; BR é o português do Brasil (pt-BR);
N-Grams:	Tipo de função da ferramenta usada, extração de *N-Grams* do corpus integral;
5x5:	tamanho do NGrams mínimo e máximo; min.=5 e max.=5
F:	frequência mínima; F=3;
R:	"Range" ou número de textos mínimo em que os *N-Grams* aprecem simultaneamente; apenas o TP foi analisado nesta etapa;
2015-10-16:	a data ano-mês-dia para facilitar a manutenção do corpus e a ordenação cronológica dos arquivos.

O tradutor analisa rapidamente esses arquivos para extrair dos resultados de A (5-*Grams*), B (4-*Grams*) e C (3-*Grams*) as palavras-chave pertinentes e as copia em um arquivo de texto de síntese, nomeado KW_BR_NGrams_2015-10-16.txt, em que "KW" designa um arquivo contendo as palavras-chave (*KeyWords* – KW). Define-se o "KW", nesse estudo, como uma palavra única ou conjunto de N palavras selecionadas após a análise dos *N-Grams* gerados pelo *AntConc* para compor a expressão de busca do mecanismo de busca (*Google Search Engine*). Essa compilação final de 20 a 30 palavras-chave será usada na etapa seguinte para a coleta de textos pertinentes por meio de um mecanismo de busca.

Eis a seguir dois exemplos de KW para cada resultado (A, B e C), com a numeração dos "Grams", em índice:

Quadro 2 – Exemplos de KWs selecionados para a coleta
Fonte: elaborado pelo autor

KW_BR_NGrams_2015-10-16.txt	N-Grams
métodos$_1$ de$_2$ elaboração$_3$ de$_4$ espumantes$_5$	5
adição$_1$ de$_2$ dióxido$_3$ de$_4$ enxofre$_5$	5
contato$_1$ com$_2$ as$_3$ leveduras$_4$	4

fermentação$_1$ em$_2$ recipiente$_3$ fechado$_4$	4
obtenção$_1$ do$_2$ mosto$_3$	3
colheita$_1$ das$_2$ uvas$_3$	3

Etapa 2A. Coletar os textos com as “palavras-chave” (KW)

A segunda etapa implementa a coleta dos textos pertinentes por meio da SEO, a partir das palavras-chave selecionadas na etapa anterior. O maior interesse dessa metodologia é, sem dúvida, a possibilidade para o tradutor de ter acesso a vários parâmetros de busca avançada para afunilar os resultados de acordo com suas necessidades. Dentro dos mecanismos de busca disponíveis na Web, *Google Search Engine* oferece o maior número de configurações acessíveis ao usuário. Vale destacar que tanto os tradutores aprendizes, quanto os profissionais podem explorar todos os recursos disponíveis.

1. As “Configurações de pesquisa”

Google Search Engine permite configurar o número de resultados máximo que possam aparecer em cada página de resultados (SERP). Em buscas orientadas para um processo de TEsC, o valor selecionado equivale, neste estudo, a 50 resultados máximo por página, pois corresponde a um limite (empírico) que gera resultados que possuem ainda uma boa aderência entre a palavra-chave (KW) inserida na caixa de busca e a temática central do TP. A figura seguinte apresenta os procedimentos para realizar essa configuração:

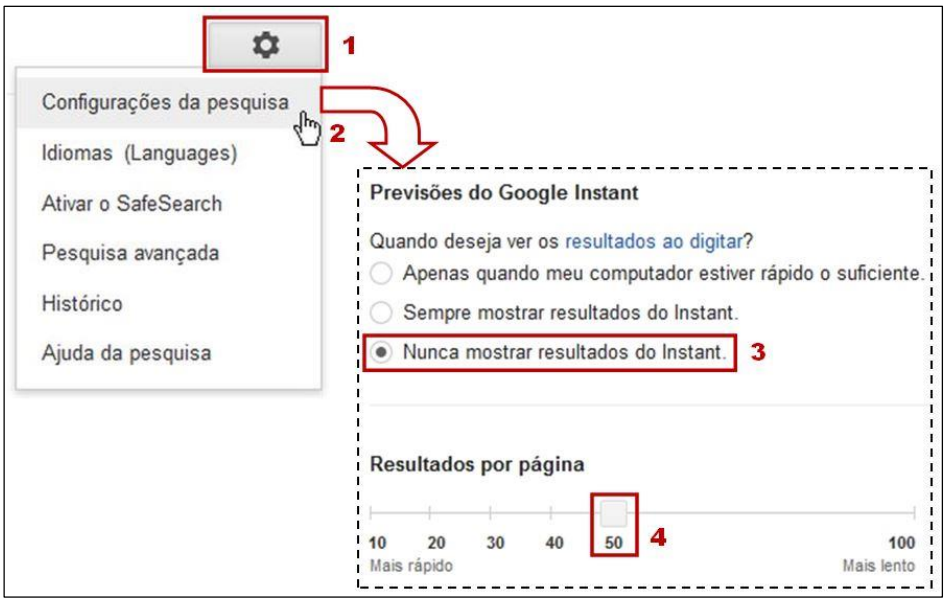


Figura 3 – Formulário das “Configurações de pesquisa” de *Google Search Engine*

Fonte: elaborado pelo autor

2. As configurações de uma “Pesquisa avançada”

A “pesquisa avançada” oferece uma “mais-valia” para a coleta de textos automatizada, pois o tradutor pode ajustar, sem conhecimentos específicos, os filtros que exibiram na página de resultados (SERP) o acesso a textos com maior pertinência.

No exemplo seguinte, o KW usado é “métodos de elaboração de espumantes”, conforme as configurações seguintes:

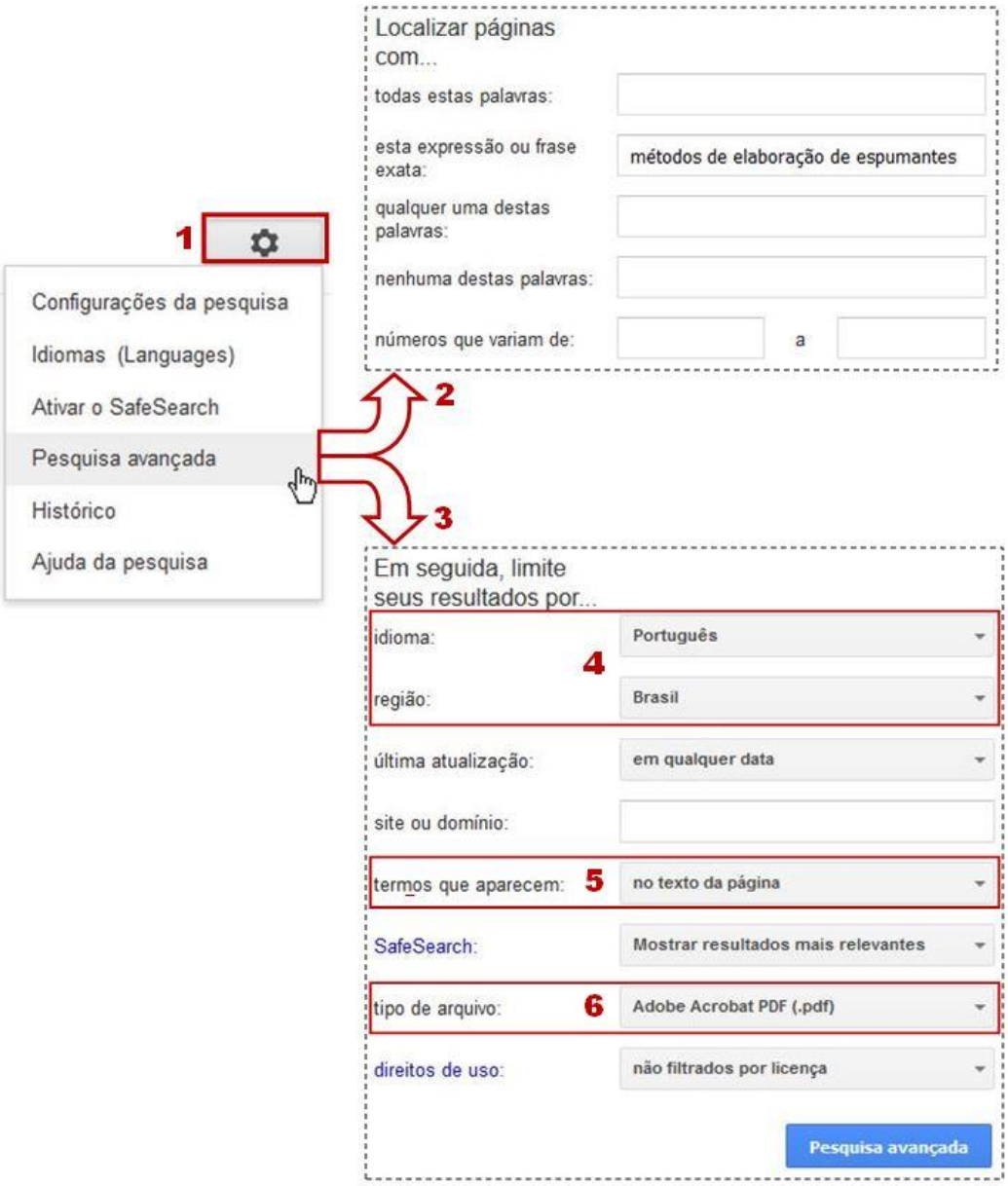


Figura 4 – Formulários de configurações da "Pesquisa avançada" de *Google Search* para a LP

Fonte: elaborado pelo autor

Na figura acima, o passo [1] permite ao tradutor iniciar as configurações detalhadas da busca filtrada. O passo [2] diz respeito à configuração da expressão de busca propriamente dita. O primeiro exemplo de KW (5-*Grams*), "métodos de elaboração de espumantes", foi inserido, para primeira análise dos resultados que serão gerados, na caixa "esta expressão ou frase exata". O tradutor pode escolher digitá-lo diretamente na caixa de busca principal do *Google Search*, colocando a expressão entre aspas duplas. O passo [3], mais específico, regula os outros critérios intrínsecos ao tipo de arquivo almejado. Assim, o passo [4] é o mais importante para selecionar não apenas a língua principal do conteúdo das páginas Web ou dos arquivos, mas filtra também as possíveis variantes linguísticas regionais ou diatópicas. Os critérios "idiomas" e "região" são configurados para distinguir, por exemplo, os conteúdos redigidos em português do Brasil e dos elaborados em português europeu.

Para o trabalho de versão escolhido (área da enologia) com o intuito de ilustrar essa metodologia, o valor padrão "em qualquer data" do critério "última atualização" foi mantido. Conforme *Google*, esse parâmetro ajuda encontrar "páginas atualizadas dentro do tempo especificado". Cinco opções são disponíveis, (1) em qualquer data; (2) nas últimas 24 horas; (3) na última semana; (4) no último mês; (5) no último ano. Vale destacar que a escolha das opções (2) e (3) se justifica para coleta de textos jornalísticos. Quanto à opção (5), ela é pertinente para a elaboração de corpus sobre as tecnologias de última geração (informática, tecnologias da informação, nanotecnologias, etc.). O critério "site ou domínio" restringe a busca em apenas um único site ou parte dele. Esse parâmetro é muito útil quando a coleta de textos é direcionada para um determinado repositório de artigos especializados, por exemplo, ou para o "Google acadêmico". Essa funcionalidade pode ser amplamente contemplada pela ferramenta *HTTrack* (www.httrack.com).

O passo [5], "termos que aparecem", indica ao *Google Search Engine* as partes específicas que estruturam os conteúdos da Web que devem ser pesquisadas para encontrar ocorrências compatíveis com a expressão de busca. O Google disponibiliza 4 opções. A primeira opção corresponde ao valor padrão "em qualquer lugar da página". Quanto à segunda , "em toda a página", ela é considerada como critério de pertinência mais elevada, pois o título representa, em geral, um resumo condensado do conteúdo do texto. No entanto, a opção "em toda a página, escolhida para este estudo, permite desconsiderar, uma parte significativa dos arquivos de tipo .PDF que foram criados a partir de documentos digitalizados ou fotografados, no formato "imagem", ou seja, que não receberam um tratamento de reconhecimento ótico de caracteres (OCR – *Optical Character Recognition*) para os tornarem editáveis. Essa "limpeza" prévia será útil na Etapa 5 seguinte, para converter, em lotes (*batch*), os arquivos .PDF em arquivos de textos .TXT, pois os conversores usados gastam mais tempo para tentar transformar, sem sucesso, esses arquivos .pdf de tipo "imagens" ou podem provocar um travamento do próprio conversor.

O passo [6] permite selecionar o tipo de formato dos arquivos a serem localizados. A escolha acima buscará encontrar texto em .pdf. Outros formatos como .DOC, .RTF (arquivos de textos editáveis) ou .XLS (planilhas eletrônicas) podem ser úteis na elaboração de um corpus de apoio à tradução (TEsC), para complementar a coleta de textos realizada para o formato.pdf. A seleção de arquivos .PS pode se revelar pertinente para a coleta de textos científicos, pois muitas produção na área das ciências exatas são divulgadas neste formato. A opção padrão "qualquer formato" retorna páginas no formato .HTML que geram muito ruído, devido ao "peritexto", ou seja, as informações marginais ao texto presente em uma página Web.

A configuração desses campos permanece durante o tempo em que a primeira página de resultados fica exibida, isto é, após fechá-la uma nova configuração dos parâmetros descrita acima será necessária. Todavia, durante a coleta de textos, o tradutor mantém a expressão de busca e muda apenas os KW entre as aspas, conforme a expressão de busca gerada com o exemplo acima: *allintext: "métodos de elaboração de espumantes" filetype:pdf*

3. Ativação da ferramenta *SEOquake* e exploração dos KW extraídos na Etapa 1

A ferramenta de análise para a otimização de mecanismos de busca usada para a coleta automatizada de textos é o *SEOquake* (www.*SEOquake*.com), isto é, uma extensão (*plug-in*) que pode ser instalada em navegadores como *Firefox, Chrome* e *Opera*.

O arquivo de texto *KW_BR_NGrams_2015-10-16.txt* contém a síntese dos KW selecionados na etapa anterior. Cada KW será inserido pelo tradutor na caixa de busca do Google Search [1] (Figura 5, abaixo) e, caso necessário, adaptados em função do número de resultados encontrados e apresentados na página de resultados (SERP). A adaptação dos KW se justifica quando o número de resultados é: (1) inferior a 10; (2) superior a 100 (valores empíricos).

No primeiro caso, o KW deve ser reavaliado. No exemplo apresentado na figura anterior, o KW "métodos de elaboração de espumantes" retornou apenas 2 resultados que devem ser preservados. A fim de flexibilizar o filtro, a posição das aspas do KW "métodos de elaboração de espumantes" pode ser modificada, sendo inseridas da seguinte forma: *métodos de "elaboração de espumantes"*. A página de resultados (SERP), com a nova expressão de busca apresenta 30 arquivos, no formato .PDF.

No segundo caso, o KW deve também ser reavaliado. Neste caso, o filtro deve ser mais restritivo. A solução adotada, neste estudo, propõe a inserção de mais um (ou, se for necessário, dois) KW do tipo 3-*Grams* (entre aspas) para localizar um número de resultados menor e de conteúdo de maior pertinência em relação à temática do TP. Alguns exemplos serão analisados no item 4, "Análise dos resultados"

A inserção do exemplo de KW, *métodos de "elaboração de espumantes"*, mantidas as configurações do *Google Search* apresentadas anteriormente, exibe a seguinte página de resultados:

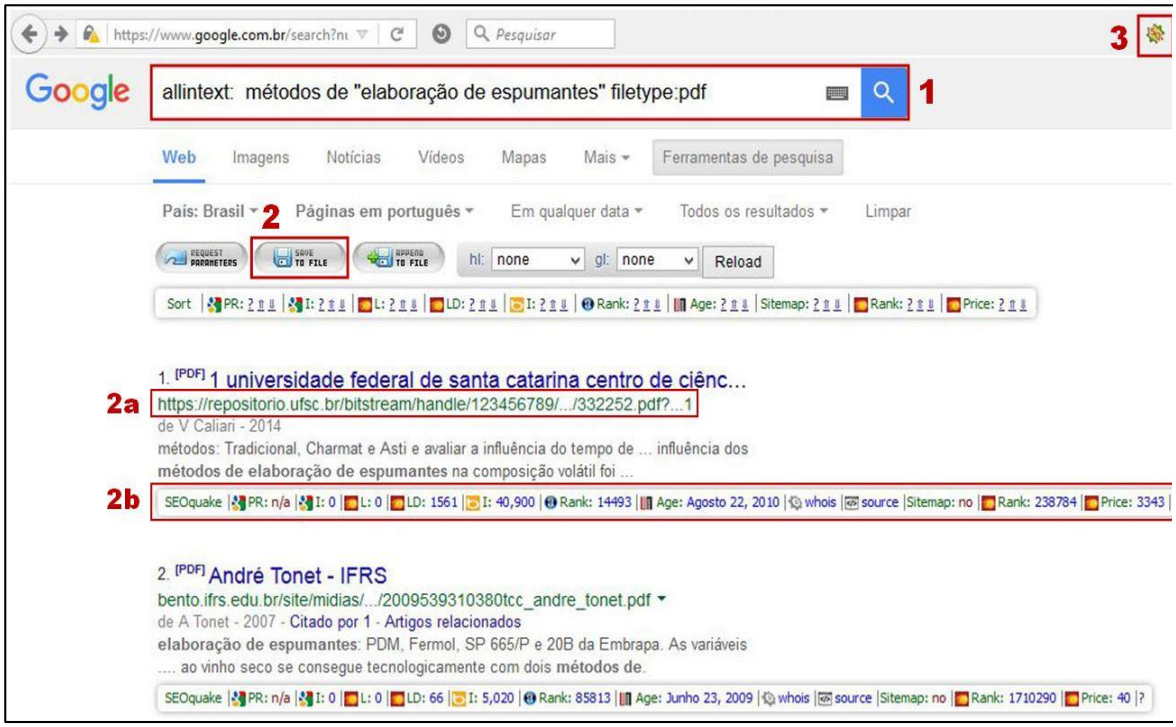


Figura 5 – Exemplo de página de resultados de busca (SERP) com *SEOquake*

Fonte: elaborado pelo autor

As funcionalidades úteis para que o tradutor possa realizar sua coleta de textos com o apoio da ferramenta de SEO são:

[1] Caixa onde a expressão de busca será digitada ou gerada automaticamente pelas configurações avançadas do *Google Search*: *allintext: métodos de "elaboração de espumantes" filetype:pdf*

[2] Comando "*Save to file*" que cria o arquivo contendo os resultados exibidos apenas na página ativa no formato .CSV, com as informações seguintes:

[2a] Primeiro link que será exportado no arquivo no formato .CSV;

[2b] Conjunto de informações SEO associadas ao link [2a] e exportadas no arquivo no formato .CSV

[3] A ferramenta *SEOquake* que pode ser desativada clicando no ícone colorido. Após desativação, esse ícone torna-se de cor cinza. Clicar novamente nele para ativar novamente a ferramenta.

Após inserção dos KW selecionados na Etapa 1 e em função dos devidos ajustamentos, o tradutor armazena os arquivos de coleta SEO, com a extensão .CSV, dentro de uma pasta nomeada, por exemplo, COLETA_CSV_BR, para os resultados na língua de partida (LP: pt-BR) e COLETA_CSV_FR para os na língua de chegada (LC: fr-FR).

Etapa 3A. Baixar em lote (batch) os arquivos selecionados (Links SEO)

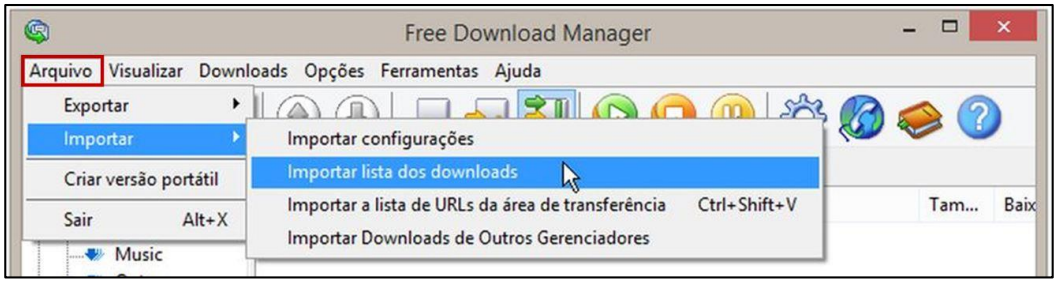
O interesse da metodologia aqui desenvolvida é proporcionar ao tradutor uma maior integração do corpus *"ad hoc"* na sua prática tradutória, automatizando alguns processos repetitivos e demorados. Essa Etapa 3 corresponde à fase de exploração dos arquivos de resultados gerados pelo *SEOquake*, isto é, baixar em lote (*batch*) os arquivos encontrados por meio de seus respectivos links. Para facilitar esse processo, deve-se proceder à unificação do conteúdo, em um único arquivo .CSV, de todos os arquivos armazenados na pasta COLETA_CSV_BR. Esse trabalho pode ser realizado por meio de um simples comando do Windows:

1. Para facilitar a digitação do caminho de acesso ao repositório "COLETA_CSV_BR", criar na raiz uma pasta: c:\coleta;
2. Copiar dentro dessa pasta (c:\coleta) todos os arquivos (.CSV e apenas eles) contido na pasta COLETA_CSV_BR;
3. Apertar as duas teclas "Windows + R", na nova janela "Executar", digitar "cmd" (sem as aspas) e clicar em "OK";
4. Digitar "cd c:\coleta" (sem as aspas) + "*Enter*";
5. Digitar "copy /b *.csv unif.csv". Todos os arquivos .CSV serão unificados em um novo e único arquivo "unif.csv" que se encontra na pasta "c:\coleta";
6. Copiar o arquivo "unif.csv" na pasta do projeto de tradução "COLETA_CSV_BR".

O arquivo "unif.csv" contém doravante todos os links encontrados conforme as expressões de busca usadas na Etapa 2. A eliminação das eventuais duplicatas de links pode ser realizada com Excel: *[Dados > Remover duplicatas]*. Em seguida, o conteúdo da coluna "Url" (do arquivo "unif.csv") deve ser copiado em um arquivo de texto .TXT, por exemplo no "Bloco de Notas" do Windows, nomeado: "download.txt". Assim, o programa gratuito de "download", *Free Download Manager* – FDM (http://www.freedownloadmanager.org/), poderá baixar automaticamente todos os arquivos .PDF disponíveis, em função de seus respectivos links presentes no arquivo "download.txt".

Para importar esse arquivo, é preciso abrir FDM, clicar em *[Arquivo > Importar > Importar lista dos downloads]*, selecionar e abrir "download.txt":

Figura 6 – Importação da lista dos links para download em lote com FDM



Fonte: elaborado pelo autor

O programa FDM precisa conhecer o caminho e o nome da pasta que vai receber os arquivos que serão baixados em lotes, na pasta de Corpus comparável correspondente de ao projeto de tradução (TEsC) e a uma das língua de trabalho, ou seja, aqui, CORPUS_ENO_BR_COMPARAVEL (ver a "arvorecência", na etapa seguinte).

Durante o processo de coleta de textos, em lotes, por meio de palavras-chave que, por vezes, são semanticamente adjacentes, podem-se encontrar vários arquivos compartilhando um conteúdo idêntico, porém com nomes diferentes e oriundos de diversos sites (links). Essa limpeza não pode ser realizada com a função "Remover duplicatas" do Excel, mas sim, com uma ferramenta, como "*Duplicate Cleaner Free*" (www.duplicatecleaner.com), capaz de comparar os conteúdos e eliminar aqueles que são idênticos. Esse rápido procedimento contribui para não confundir o tradutor no seu processo decisório (TEsC), com estatísticas de ocorrências errôneas (devido aos conteúdos idênticos), geradas pelo concordanciador *AntConc*.

Etapa 4A. Organização do Corpus Comparável na LP

O texto de partida, documento de apoio para nosso estudo, é um texto científico da área da enologia (ENO), "Vinhos espumantes: métodos de elaboração" (CALIARI et al., 2013), que deve ser traduzido (versão) do português do Brasil (pt-BR) para o francês (fr-FR). A área de conhecimento é um critério básico (três letras em caixa alta) para a nomeação das pastas e dos arquivos que constituem os corpora, organizados conforme a estrutura em arvorecência seguinte:

CORPORA_ENO
└─ 01_CORPUS_ENO_BR_COMPARAVEL
└─ 02_CORPUS_ENO_BR_CUSTOMIZADO

Após criar essas pastas, o tradutor copia o conjunto de textos armazenados 01_CORPUS_ENO_BR_COMPARAVEL. A renomeação dos arquivos .PDF é realizada automaticamente, em lotes (*batch*), com a ferramenta "*Bulk Rename Utility*" (www.bulkrenameutility.co.uk). Os novos nomes *KW_ENO_BR_Ngram_2015-10-001.pdf* seguem a regras seguinte:

KW:	arquivo encontrado por meio de uma determinada palavra-chave (KW) ou conjunto de palavras-chave;
ENO:	as três letras que identificam a área de conhecimento do corpus;
BR:	a língua do arquivo .PDF;
N-gram:	o grupo de KW que conduziram o mecanismo de busca a encontrar esse arquivo .PDF.
2015-10-16:	ano-mês-dia da criação dessa lista de KW;
-001:	número de ordem automático do arquivo .PDF.

Vale destacar que no processo de renomeação não há: (1) espaços, sempre preenchido por _ (*underscore*); (2) sinais diacríticos (ã, Ã, ç, Ç, etc.), substituídos por letras simples (A, C, etc.). Essas precauções se justificam quando os corpora são enviados por e-mail, armazenados nas nuvens ou em um servidor FTP, por exemplo, os espaços podem ser substituídos por "%20", Ã por "%C3" ou "Ã§"; Ç por "%C7" ou "Ã£", tornando os nomes quase ilegíveis. Exemplo de renomeação (excerto) obtido com a ferramenta "*Bulk Rename Utility*":

Nomes originais dos arquivos baixados	Arquivos renomeados
305f30be869020d4adb7078ce30d4091.pdf	'KW_ENO_BR_Ngram_2015-10-001.pdf
00000I70U.PDF	KW_ENO_BR_Ngram_2015-10-002.pdf
04_Bebidas_Fermentadas.pdf	KW_ENO_BR_Ngram_2015-10-003.pdf

Quadro 3 – Exemplo de renomeação (excerto) obtido com a ferramenta "*Bulk Rename Utility*"
Fonte: elaborado pelo autor

Os arquivos presentes na pasta 01_CORPUS_ENO_BR_COMPARAVEL devem ser renomeados antes de serem convertidos no formato .TXT, a fim de estabelecer um vínculo entre o arquivos original, com a extensão .PDF e o arquivo convertido para o formato .TXT. Essa ligação é importante para agilizar o saneamento de dúvidas que poderiam surgir durante o processo de análise dos corpora com *AntConc*, por meio de uma rápida consulta do arquivo original .PDF.

Etapa 5A. Organização do Corpus Customizado na LP

Os corpora customizados (Perrotti-Garcia, 2005) são aqueles que serão explorados diretamente pelo tradutor durante o processo de *Tradução embasada em Corpora* (TEsC), com um concordanciador, como *AntConc*. Essa ferramenta reconhece apenas arquivos na extensão .TXT. Portanto, o corpus comparável elaborado na etapa anterior deve ser convertido, em lotes (*batch*), para esse formato de trabalho. O conversor de arquivos .PDF para .TXT, "*Some pdf to txt*" (www.somepdf.com), pode realizar essa tarefa. Porém, vale ressaltar que, na sua configuração padrão, os arquivos são codificados em ANSI (Latin1), por isso, será necessário modificar a configuração padrão ("*Text encoding*") do concordanciador *AntConc* (v.3.4.4) de UFT8 para ANSI ou do próprio conversor de ANSI (Latin1) para UFT8.

Todavia, Laurence Anthony, o autor de *AntConc* , propõe uma ferramenta gratuita, "*AntFileConverter*" (www.laurenceanthony.net), que converte os arquivos .PDF ou .DOCX para a extensão .TXT codificados em UTF8, plenamente compatível com o concordanciador *AntConc* . Caso seja necessário converter a codificação ANSI de arquivos .TXT para UTF8, a ferramenta "EncodeAnt[5]" detecta a codificação de um arquivo ou de um corpus e pode recodificá-lo, em lotes, para UTF8.

Para evitar o manuseio de um grande número de documentos, os arquivos .PDF são convertidos diretamente na pasta 01_CORPUS_ENO_BR_COMPARAVEL e, após a conclusão do processo, os arquivos convertidos .TXT, com os mesmos nomes dos seus respectivos arquivos .PDF de origem, são recortados e copiados na pasta de trabalho final 02_CORPUS_ENO_BR_CUSTOMIZADO. O conversor "*Some pdf to txt*" oferece a possibilidade de escolher a pasta de destinação dos arquivos convertidos ("*Output Directory*").

[5] Também da autoria de Laurence Anthony (www.laurenceanthony.net)

Devido a esses cuidados referentes à codificação, aconselha-se de mencionar seu tipo na hora de nomear os arquivos .TXT, para evitar elaborar corpora customizados compostos de textos codificados de forma diferente. O concordanciador só pode ler apenas um tipo ao mesmo tempo, as letras que possuem sinais diacríticos serão alteradas e, por conseguinte, desconsideras na pesquisa de concordância. Por exemplo, a palavra de busca "canção", pode ser exibida como "can��o" ou "canÃ§Ã£o".

A sintaxe dos nomes dos arquivos que compõem os corpora customizados pode acrescentar o tipo de codificação conforme as necessidades, com a ferramenta "*Bulk Rename Utility*", antes do número de ordem (sempre no final do nome):

KW_ENO_BR_Ngram_2015-10-ANSI-001.txt [ou]
KW_ENO_BR_Ngram_2015-10-UTF8-001.txt

Essa metodologia enfatiza a importância do papel da sintaxe dos nomes de arquivos e de pastas, pois o tradutor deve ser capaz identificar o máximo de características dos arquivos (tipo, origem, data, codificação, etc.) em 20 ou 30 caracteres.

O primeiro corpus customizado (LP) está disponível para ser explorado.

Fase B – Elaboração dos corpora na LC

A fase B do processo de elaboração de corpora na LC para a *Tradução embasada em Corpora* (TEsC) constitui-se das mesmas etapas e procedimentos da fase A que se focaliza apenas na LP (ver o Quadro 1 e a Figura 1, acima). No entanto, vale destacar que apenas a Etapa 1B distingue-se da Etapa 1A, no que diz respeito à lista de KWs usados na coleta automatizada dos textos na LC. As demais etapas da Fase B seguem o mesmo método de realização e de desempenho da Fase A.

Etapa 1B. Elaborar uma lista de "palavras-chave" na LC

A elaboração da lista de "palavras-chave" na LC fundamenta-se diretamente no resultado da extração terminológica realizada na Etapa 1A. Os KWs da LC são gerados por meio de "hipóteses de tradução". Essa tradução justifica-se, nessa metodologia, para garantir, de forma rápida e confiável, certa aderência dos textos que serão coletados com a temática central do TP. O contexto imediato dos KWs não está mais presente no arquivo *KW_BR_NGrams_2015-10-16.txt*, criado na Etapa 1A, mas sim, de modo geral, conhecido pelo tradutor que constrói seus próprios corpora.

Assim, a pertinência da tradução desses KWOC (*Keyword out of Context*) será diretamente avaliada no mecanismo de busca (*Google Search Engine*) antes de ser aprovado para compor a lista de KW na LC, armazenados no arquivo *KW_BR-FR_**HIP**_NGrams_2015-10.txt.* Nessa etapa, é também possível abrir uma seção da ferramenta *AntConc*, com apenas o TP para verificar o contexto de determinados KWs (KWIC – *Keyword in Context*) e melhorar as hipóteses de tradução ou sanar eventuais ambiguidades.

Em nosso estudo, os idiomas de trabalho pertencem à mesma família linguística, isto é, as das línguas românicas ou neolatinas (português e francês). Neste contexto, o tradutor pode apoiar-se no conceito de "tradução transparente" (Miroir, 2014) para emitir suas hipóteses de tradução. A tradução transparente é uma tradução literal que se integra plenamente na idiomaticidade da língua de chegada. Assim sendo, os exemplos seguintes mostram resultados concretos da aplicação desse método:

Quadro 4 – Exemplos de hipóteses de tradução com suas respectivas adequações
Fonte: elaborado pelo autor

	KW_BR_NGrams_2015-10-16.txt	KW_BR-FR_HIP_NGrams_2015-10.txt
1	adição de dióxido de enxofre	**addition | ajout** de dioxyde de soufre
2	métodos de elaboração de espumantes	méthodes d'élaboration de mousseux | vins effervescents
3	fermentação em recipiente fechado	fermentation ~~en récipient fermé~~ ➔ **fermentation en cuve close**
4	grupo das moscatéis	~~groupe~~ des muscats ➔ **famille des muscats**

No exemplo nº1, Quadro 4, a tradução transparente de "adição de dióxido de enxofre" para o francês gera o grupo nominal "*addition de dioxyde de soufre*". Ao verificar a pertinência da proposta o tradutor percebe que o mecanismo de busca fornece outras propostas pertinentes (KWIC) como "*ajout de dioxyde de soufre*" com frequências de uso bastante equivalentes com o KW inicial. A palavra "*vins*" (vinhos) é um marcador de coerência contextual.

Figura 7 – Exemplos de KWIC (negrito) gerados por *Google Search Engine*

Fonte: elaborado pelo autor

Neste caso, o KW mantido, no arquivo *KW_BR-FR_HIP_NGrams_2015-10.txt*, será "***addition* | *ajout*** *de dioxyde de soufre*", com a barra vertical equivalente à função booleana "OR": as palavras "*addition*" ou "*ajout*", usadas simultaneamente na mesma expressão de busca. Observa-se o mesmo fenômeno linguístico no exemplo nº2, a tradução transparente de "métodos de elaboração de espumantes" corresponde ao conjunto "*méthodes d'élaboration de mousseux*", em que "*mousseux*" aparece frequentemente como equivalente de "*vins effervescents*": KW gravado no arquivo *KW_BR-FR_HIP_NGrams_2015-10.txt*, "*méthodes d'élaboration de* ***mousseux* | *vins effervescents***".

No exemplo nº3, a hipótese elaborada na base de uma tradução transparente não foi satisfatória, pois não gerou resultados relevantes. Assim a versão da palavra-chave "fermentação em recipiente fechado" (com a presença das aspas) para "*fermentation en récipient fermé*", não pose ser considera "transparente", mas sim, apenas literal, no contexto do TP. Há, neste caso, necessidade de reformular a hipótese de tradução para se adequar às exigências da língua de chegada. O mecanismo de busca gerou resultados terminológicos pertinentes como "*fermentation en cuve close*" que, após verificação, confirmou-se a relevância do KW no contexto terminológico do TP, foi arquivado em *KW_BR-FR_HIP_NGrams_2015-10.txt.* O mesmo procedimento foi aplicado para resolver a geração de KW na LC, no exemplo nº4: "grupo das moscatéis" pode ser traduzido por "*groupe des muscats*", porém, após o teste de pertinência, os resultados não foram convincentes. A adaptação do sintagma de "*groupe des muscats*" para "*famille des muscats*", solução proposta pelo mecanismo de busca, permitiu acessar a documentos com conteúdo similar ao do TP. O KW foi arquivado em *KW_BR-FR_HIP_NGrams_2015-10.txt.*

Essa metodologia de provas e sucessivas aproximações levam o tradutor a selecionar KWOC (Palavras-chave fora do contexto) na LC, mais próximos da temática central do texto (ou conjunto de textos) a traduzir. O tradutor, na base do arquivo *KW_BR-FR_HIP_NGrams_2015-10.txt*, por ele elaboro, pode iniciar seu processo de coleta de textos na LC, seguindo as mesmas etapas da metodologia apresentada anteriormente para a elaboração dos corpora na LP.

Etapa 2B. Coletar os textos com as "palavras-chave" (KW)

O tradutor reitera as mesmas ações das aquelas descritas na Etapa 2, observando que as configurações avançadas de língua e região do *Google Search* devem ser modificadas em função das características exigidas pelo TC [4]. Os demais critérios [5] e [6] permanecem idênticos:

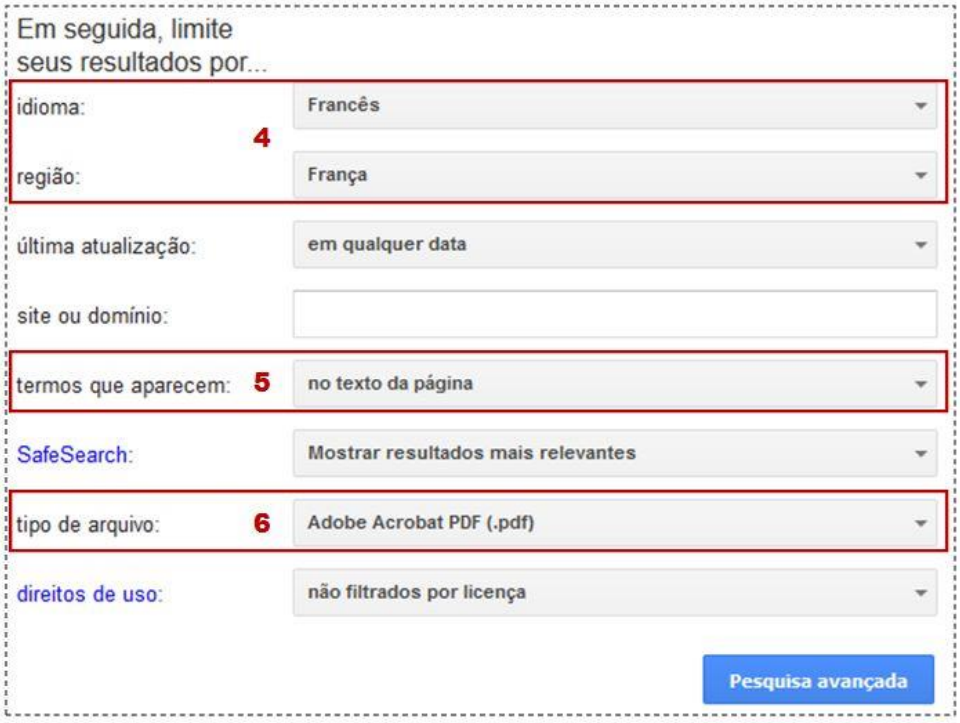


Figura 8 – Formulários de configurações da "Pesquisa avançada" de *Google Search* para a LC

Fonte: elaborado pelo autor

Conforme os procedimentos definidos na Etapa 2A, o tradutor aplica a mesma metodologia. Na base dos KW definidos e selecionados na Etapa 1B, ele armazena os arquivos da coleta SEO (com a extensão .CSV), na pasta nomeada COLETA_CSV_FR.

Etapa 3B. Baixar em lote (batch) os arquivos selecionados (Links SEO)

Da pasta COLETA_CSV_FR, foi criado, como na Etapa 3A, o arquivo unificado "unif.csv" que contém todos os links encontrados conforme as expressões de busca usadas na Etapa 2B. O programa *Free Download Manager* – FDM, pode baixar automaticamente todos os arquivos .PDF disponíveis seguindo os links presentes no arquivo "download.txt".

Etapa 4B. Organização do Corpus Comparável na LC

O tradutor cria na pasta já existente CORPORA_ENO, com as subpastas seguintes:

CORPORA_ENO
└─ 01_CORPUS_ENO_FR_COMPARAVEL
└─ 02_CORPUS_ENO_FR_CUSTOMIZADO

A renomeação dos arquivos .PDF é realizada automaticamente, em lotes (*batch*), com a ferramenta "*Bulk Rename Utility*". Os novos nomes *KW_ENO_FR_Ngram_2015-10-001.pdf* seguem a sintaxe exposta na Etapa 4A.

Etapa 5B. Organização do Corpus Customizado na LC

O conversor de arquivos .PDF para .TXT, "*Some pdf to txt*", realiza essa tarefa. Porém, vale ressaltar que os arquivos devem ser codificados em UFT8 para padronizar os arquivos que serão explorados com o concordanciador *AntConc*. Os arquivos podem ser renomeados com a menção de codificação: *KW_ENO_FR_Ngram_2015-10-**UTF8**-001.txt*

O segundo corpus customizado (na LC) está disponível para ser explorado, juntamente com o primeiro corpus customizado (na LP) elaborado na Fase A.

Fase C – Implementação do processo de *Tradução embasada em Corpora* (TEsC)

Por questões de espaço, a Fase C não será analisada detalhadamente no âmbito desse trabalho. No entanto, a Figura 9 ilustra os procedimentos da TEsC com os dois corpora elaborados nas Fases A e B e um editor de tradução, como o *Wordfast Pro* (www.wordfast.com).

O tradutor abre simultaneamente duas seções do *AntConc* e carrega o corpus customizado da LP (lado direito de sua tela) e o da LC (lado esquerdo de sua tela). Durante o processo tradutório, com (ou sem) o editor de tradução, ele pode consultá-los indiferentemente para, por exemplo, testar suas hipóteses de tradução e verificar a coerência terminológica tanto na LP quanto na LC. Durante essa fase C, a ferramenta "*Concordance*" do concordanciador *AntConc* será de grande utilidade.

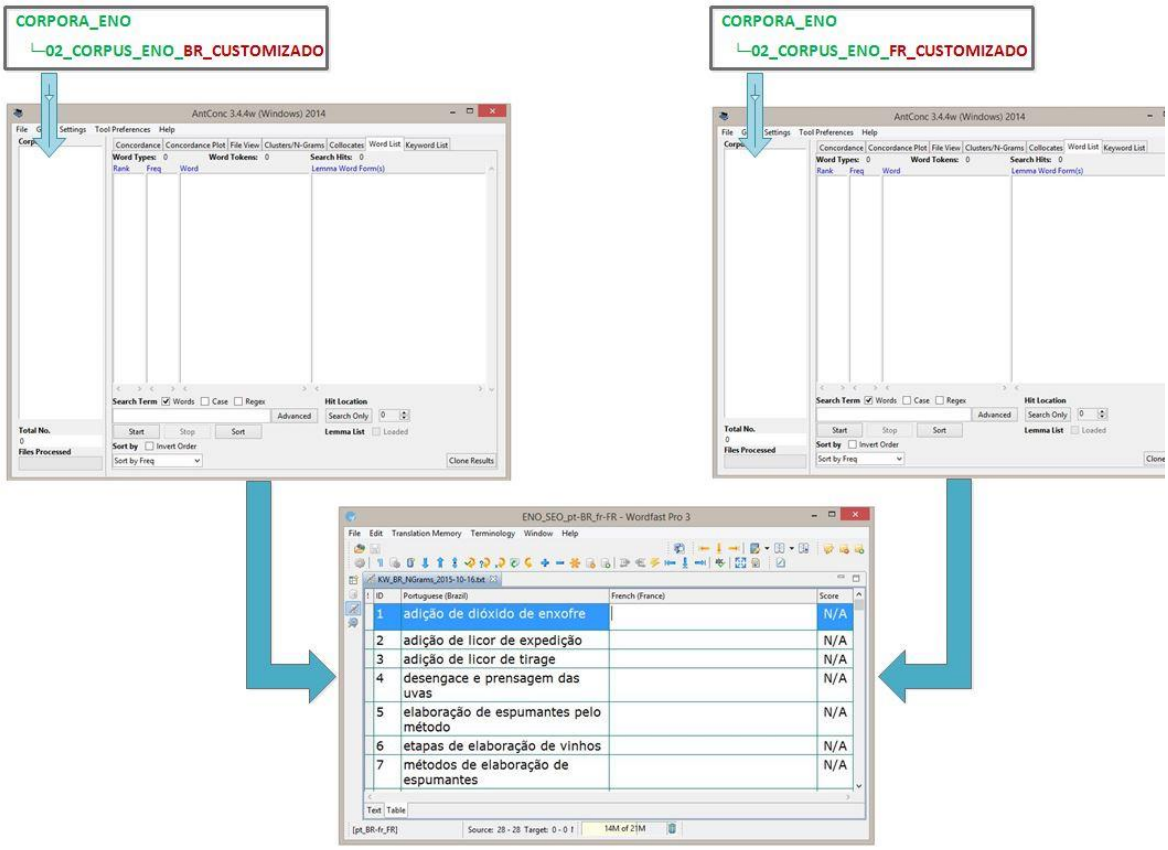


Figura 9 – Processo de *Tradução embasada em Corpora* (TEsC)

Fonte: elaborado pelo autor

4. Análise dos resultados

Fase A – Elaboração dos corpora na LP

Etapa 1A. Extração das "palavras-chave" do TP

O processo de extração das palavras-chave do TP, composto, após uma limpeza básica, de 1109 *types* (nº de palavras únicas) e de 4511 *tokens* (nº de ocorrências de todas as palavras), gerou um conjunto de 39 palavras-chave (KW) significativos para iniciar a coleta de textos, ou seja, 21,2 % do total dos *N-Grams* extraídos. A tabela seguinte apresenta a síntese dos resultados obtidos, com A (5-*Grams*), B (4-*Grams*) e C (3-*Grams* ou tri-*Grams*):

Tabela 1 – Síntese dos resultados de extração dos KWs do TP
Fonte: elaborado pelo autor

	A	B	C	Total
N-Grams size: min.	5	4	3	
N-Grams size: max.	5	4	3	
Min. Freq.	3	3	3	
Min. Range	1	1	1	
Nº de *N-Grams* obtidos	31	52	101	184
Nº de *N-Grams* válidos	3	7	29	39

Razão N-Gr. val./obt.	9,7%	13,5%	28,7%	21,2%

A razão entre os *N-Grams* válidos e os obtidos parece, em primeira análise, muito baixa para A (9,7%). Ao iniciar seu processo de extração de KW pelo arquivo *AntConc_BR_N-Grams_5x5_F3_R1_2015-10-16.txt*, o tradutor, na sua breve análise do conteúdo, vai observar a sequência seguinte de 5-*Grams* e proceder a uma reorganização eliminando os 5-*Grams* incompletos ou de pouco relevância para se tornarem um KW. No exemplo seguinte, os núcleos dos KWs são as palavras "elaboração" e "espumantes":

Quadro 5 – Exemplos de análise para a extração dos KWs do TP (5-*Grams*)
Fonte: elaborado pelo autor

N-Grams_5x5	Após análise ...	KW
métodos de elaboração de espumantes	métodos de elaboração de espumantes	5-Grams
de elaboração de vinhos espumantes	elaboração de vinhos espumantes	4-Grams
elaboração de espumantes pelo método	elaboração de espumantes	3-Grams
a elaboração de espumantes pelo	elaboração de espumantes	repetição
para a elaboração de espumantes	elaboração de espumantes	repetição
para a elaboração de vinho	elaboração de vinho	3-Grams

Observa-se, no quadro acima, que após a remoção dos artigos, das preposições ou incompletudes ("pelo método..."), esse conjunto de 6 ocorrências de 5-*Grams*, transformou-se em apenas quatro KWs: um 5-*Grams*, um 4-*Grams* e dois 3-*Grams*. Os dois 5-*Grams* iniciais, "a elaboração de espumantes pelo" e "para a elaboração de espumantes", foram eliminados devido ao efeito de repetição com o novo 3-*Grams* "elaboração de espumantes". No entanto, o tradutor pode julgar pertinente, para determinados casos, manter o KW com a preposição "para a elaboração de espumantes", com intuito de introduzir uma restrição maior em função do número de resultados de busca exibidos. Aconselha-se, nesse caso, o acréscimo na expressão de busca de um 3-*Grams* ou 2-*Grams* pertinentes.

Percebe-se que a mesma estruturação de KW, do que a realizada anteriormente (Quadro 5), pode ser aplicada na análise dos outros dois arquivos formados por 3-*Grams* e 4-*Grams*. Assim, as ocorrências, como "elaboração de vinhos espumantes" e "a elaboração de espumantes", que constam no arquivo *AntConc_BR_N-Grams_4x4_F3_R1_2015-10-16.txt,* são rapidamente desconsideradas pelo tradutor, pois já foram selecionadas na análise anterior do arquivo de 5-*Grams*. A avaliação dos 4-*Grams* seguintes, "de dióxido de enxofre" e "adição de dióxido de", podem ser escolhidos como dois 3-*Grams* "dióxido de enxofre" e "adição de dióxido", mas também gerar um 5-*Grams* pertinente: "adição de dióxido de enxofre".

Destaca-se, no arquivo *AntConc_BR_N-Grams_3x3_F3_R1_2015-10-16.txt,* o mesmo fenômeno, 3-*Grams*, "segunda fermentação em"; "uma segunda fermentação", tornaram-se 2-*Grams* pertinentes para a temática principal do TP: "segunda fermentação". A lista de 3-*Grams* (ou tri-*Grams*) é composta de 31 ocorrências significativas e variadas e gerou um conjunto de quatro 2-*Grams* (ou bi-*Grams*), por exemplo:

Quadro 6 – Exemplos de análise para a extração dos KWs do TP (3-*Grams*)
Fonte: elaborado pelo autor

Após análise ...	Tipo de KW
envelhecimento em contato fermentação em recipiente grupo das moscatéis tomada de espuma [...]	3-Grams
método charmat segunda fermentação vinhos espumantes vinhos tranquilos	2-Grams

A elaboração da lista de palavras-chave é, na prática, bastante rápida, pois o tradutor salva apenas os *N-Grams* pertinentes, ou parte deles, no arquivo *KW_BR_N-Grams_2015-10-16.txt*. O resultado pode ser considerado satisfatório e suficiente, pois as palavras-chave serão associadas entre si, de forma combinatória em função do número de resultados de busca exibidos na página (SERP). No que se refere ao robô de coleta automatizada de textos, "*BootCat ToolKit*" (www.bootcat.sslmit.unibo.it), os autores nomeiam "seed", o KW e "tuple" um conjunto de "seeds" ou grupo de KWs, que compõem a expressão de busca.

Etapa 2A. Coleta dos textos com as "palavras-chave" (KW)

Após a devida configuração dos critérios do mecanismo de busca *Google Search* e a ativação do *SEOquake*, o tradutor explora a lista de KW que estabeleceu na etapa anterior.

Nesse estudo, observou-se que o *SEOquake* possui o melhor desempenho com o navegador *Firefox*. Por exemplo, os arquivos de resultados .CSV, gerados pelos *SEOquake* no *Chrome*, são sempre nomeados da mesma forma, isto é, "data.csv", independentemente da expressão de busca inserida. O *Firefox*, por sua vez, nomeia automaticamente os arquivos gerados com a expressão de busca completa. A tabela seguinte apresenta as expressões de busca geradas pelo *Google Search* e os nomes dos arquivos .CSV gerado automaticamente pelo *SEOquake* no *Firefox*:

KW extraído com posição das aspas (filtragem)	Expressão de busca no Google Search ([1] na figura acima)	Nome do arquivo .CSV gerado no Firefox
"métodos de elaboração de espumantes"	allintext: métodos de "elaboração de espumantes" filetype:pdf	allintext- -métodos de elaboração de espumantes-filetype-pdf.csv
métodos de "elaboração de espumantes"	allintext: métodos de "elaboração de espumantes" filetype:pdf	allintext- métodos de - elaboração de espumantes-filetype-pdf.csv

Quadro 7 – Exemplos expressões de busca geradas pelo *Google Search* e seus respectivos arquivos .CSV do Firefox

Fonte: elaborado pelo autor

Conforme o Quadro 7, o nome do arquivo .CSV, criado no Firefox, contém as informações seguintes:

- allintext- || critério do *Google Search* avançado, "termos que aparecem" com a opção "Pesquise por termos em toda a página";

- métodos de -elaboração de espumantes- || KW com a posição das aspas assinaladas pelos hifens;
- filetype-pdf || tipo de arquivo selecionado;
- .CSV || formato do arquivo de resultados do *SEOquake*.

Uma pasta de coleta de textos, com o *SEOquake*, apresenta-se da seguinte forma, com seus arquivos (excerto) nomeados automaticamente, em função dos dados de configuração do Google Search avançado e da expressão de busca inserida:

COLETA_CSV_BR

[1] └─ allintext- -adição de leveduras- filetype-pdf (55 links)

[2] └─ allintext- -adição de leveduras- -vinhos- filetype-pdf (36 links)

[3] └─ allintext- -adição de leveduras- -vinhos espumantes- filetype-pdf (11 links)

O tradutor pode administrar suas pastas de coletas (COLETA_CSV_BR; COLETA_CSV_FR) e conduzir sua pesquisa com maior rapidez, conforto e segurança, no que se refere à gestão dos resultados da análise SEO. Percebe-se, no exemplo acima, que a leitura dos nomes dos arquivos demonstra o processo de elaboração do corpus. Assim, o KW "adição de leveduras" [1], gerou 55 resultados (links para o SEO). O tradutor resolveu restringir o filtro com uma palavra pertencendo ao contexto e à temática do TP a traduzir: "vinhos" [2]. Essa expressão de busca retornou 36 links compatíveis. O filtro foi mais restritivo com o acréscimo, na mesma filosofia de trabalho, de "vinhos espumantes" [3] e apresentou 11 resultados. Essa metodologia não apresenta risco de acumular vários arquivos idênticos, pois as duplicatas serão removidas na etapa seguinte.

No âmbito desse estudo, foram coletados 30 arquivos de resultados .CSV, na base de 30 expressões de busca (EB) distintas: 14 com KW únicos; 13 com 2 KW associados (função AND); 3 com 3 KW associados (função AND):

Nº de EB	KW1	[AND] KW2	[AND] KW3
14	adição de dióxido de enxofre		
	adição de leveduras selecionadas		
	autólise das leveduras		
	contato com as leveduras		
	[...]		
13	adição de dióxido	adição de leveduras	
	colheita das uvas	com as leveduras	
	etapas de	elaboração de vinhos espumantes	
	rolhamento	engaiolamento	
	obtenção do mosto	espumante|espumantes	
	[...]		
3	rolhamento	engaiolamento	embalagem
	dióxido de carbono	dióxido de enxofre	espumante
	dióxido de carbono	dióxido de enxofre	espumante|espumantes

Quadro 8 – Exemplos de "expressões de busca" com associações de KWs

Fonte: elaborado pelo autor

Etapa 3A. Transferência (download) em lote (batch) dos arquivos selecionados (Links SEO)

O conteúdo desses arquivos, com a extensão .CSV, é facilmente explorável com um programa de planilhas eletrônicas (*Microsoft-Excel* ou *LibreOffice-Calc*) e apresenta-se da seguinte maneira, como para o arquivo: *allintext- métodos de -elaboração de espumantes-filetype-pdf.csv* (excerto):

Url	Google PageRank
http://r1.ufrrj.br/cpda/wp-content/uploads/2012/07/Tese_Paulo_Nierdele_20111.pdf	1
http://www.cnpuv.embrapa.br/publica/anais/cbve12/cbve12.pdf	1
http://www.cnpuv.embrapa.br/publica/documentos/doc048.pdf	1
http://www.scielo.br/pdf/cta/v31n1/28.pdf	1
http://ecoinovar.com.br/cd2013/arquivos/artigos/ECO229.pdf	0
http://pitangui.uepg.br/departamentos/defito/labiovegetal/Viticultura.pdf	0
https://repositorio.ucs.br/jspui/bitstream/1131/Dissertacao%20Laercio%20Spadari.pdf	n/a

Tabela 2 – Exemplos de "Url" gerados com *SEOquake* e seus respectivos PageRank

Fonte: elaborado pelo autor

A coluna "Url" é a mais importante para o tradutor, pois esses links importados da página de resultados (SERP) do *Google Search* serão usados, nessa etapa, para baixar, automaticamente e em lotes (*batch*), os arquivos de seus respectivos sites e/ou repositórios. Quanto à coluna "*Google PageRank*", ela exibe os dados que indicam a relevância da página (ou do site) para os algoritmos do Google. Essa métrica varia em uma escala de 0 a 10 (site "nota 10" para Google), a abreviação "n/a" significa "*not available*", ou seja, o *PageRank* (PR) não disponível para essa página. Assim, o tradutor pode classificar os arquivos pelo PR e baixar, por exemplo, apenas aqueles que possuem um PR superior ou igual a 0, descartando assim os arquivos com PR = "n/a".

O conjunto de 30 arquivos de resultados .CSV, coletado na etapa anterior, gerou, após o processo de unificação, o arquivo "unif.csv" compostos de 497 links. Depois da eliminação das duplicatas, com a função "Remover duplicatas" do programa Excel, 374 links únicos foram mantidos, ou seja, 75% (123 links removidos). Cada expressão de busca rendeu em média 12,5 links.

O programa de "download", *Free Download Manager* (FDM), conseguiu baixar, na pasta 01_CORPUS_ENO_BR_COMPARAVEL, 357 arquivos .PDF, a partir da lista de links presentes no arquivo "download.txt". Apenas 17 arquivos (ou seja, 5%) não foram baixados, pois, conforme as mensagens dos "logs" do FDM, os "arquivos não foram encontrados no servidor". Essas informações foram verificadas manual e individualmente. De fato, neste caso, nenhum arquivo era acessível. Esse procedimento é importante, pois alguns tipos de site, como os de repositório de textos, podem impedir o acesso a determinados arquivos por meio de uma ferramenta automatizada (robô).

A "limpeza" realizada com o programa "*Duplicate Cleaner Free*", ferramenta capaz de comparar e de eliminar os arquivos com conteúdos idênticos (duplicatas). O programa não detectou arquivos duplicados. Nessa etapa, nosso corpus coletado é constituído de 357 arquivos.

Etapa 4A. Organização dos Corpora Comparáveis

Os arquivos .PDF que constam doravante na pasta 01_CORPUS_ENO_BR_COMPARAVEL devem ser renomeados antes de serem convertidos no formato .TXT. Esse processo é realizado em lote, com a ferramenta "*Bulk Rename Utility*", e sem nenhum problema. Todos os arquivos foram renomeados conforme a "sintaxe" apresentada anteriormente: KW_ENO_BR_Ngram_2015-10-001.pdf. A operação não apresenta dificuldades.

Para evitar problemas na etapa de conversão seguinte, os arquivos são analisados com o programa de análise de metadata de arquivos PDF, Edit PDF Metadata (http://www.evermap.com). Os arquivos protegidos, por meio de uma senha, para impedir a cópia do conteúdo são eliminados. Foram encontrados 35 arquivos (em geral, teses e dissertações), ou seja, 10 % do corpus verificado.

Etapa 5A. Organização dos Corpora Customizados

A última etapa a ser realizada antes da exploração dos corpora consiste na conversão dos arquivos no formato .PDF para o .TXT para que possam ser lidos com o concordanciador usado no processo de Tradução embasado em Corpora (TEsC). Essa conversão será realizada com o programa "*Some pdf to txt*" oferece a possibilidade de escolher a pasta de destinação dos arquivos convertidos ("*Output Directory*"), isto é, 02_CORPUS_ENO_BR_CUSTOMIZADO

Conforme apresentado anteriormente na metodologia, os arquivos convertidos, que se encontram na pasta 02_CORPUS_ENO_BR_CUSTOMIZADO, pode ser renomeados, em lotes, para inserir o tipo de codificação escolhido do arquivo .TXT, de preferência UTF8: KW_ENO_BR_Ngram_2015-10-**UTF8**-001.txt

Neste estudo, foram encontrados 12 arquivos "PDF-Imagens", sem tratamento OCR, que geram arquivos .TXT vazios (0 KB), ou seja, 4% do corpus inicial de 319 arquivos a ser convertido. A versão final do corpus customizado na LC é composta de 307 arquivos disponíveis para pesquisa.

A Tabela 3 apresenta a síntese dos resultados finais do processo de elaboração dos corpora na LP, organizados por atividades específicas e com suas respectivas razões relativas (% rel.) e absolutas (% abs.). Percebe-se que, após concluir as várias etapas do procedimento, 38 % dos arquivos inicialmente coletados foram retirados do corpus. A remoção dos duplicados é a mais significativa, pois na coleta automatizada, expressões de busca diferentes podem selecionar caminhos de acesso a documentos (links) com conteúdo idênticos, devido à aderência temática do KWs selecionados.

Descrição atividade	Nº item	Tipo	% rel.	% abs.	Observações
Coleta SEO "unif.csv"	497	links			
Remover duplicatas	374	links	-25%	-25%	123 links duplicados
Download concluídos	357	.pdf	-5%	-28%	13 downloads impossíveis; 4 arquivos HTML
Conteúdo duplicado	357	.pdf	0%	-28%	0 encontrados
Edit PDF Metadata	322	.pdf	-10%	-35%	35 arquivos protegidos
Some PDF to TXT	319	.txt	-1%	-36%	3 arquivos inválidos
Nº de textos válidos	307	.txt	-4%	-38%	12 arquivos vazios (sem OCR)

Tabela 3 – Síntese final do processo de elaboração dos corpora na LP

Fonte: elaborado pelo autor

A análise do corpus 02_CORPUS_ENO_BR_CUSTOMIZADO (sem limpeza), por meio da ferramenta "Word List" do concordanciador *AntConc*, apresentou os resultados seguintes: 125.885 *Word Types* e 5.407.726 *Tokens*.

O corpus elaborado na língua de partida (LP) está pronto para ser explorado com a concordanciador *AntConc*, no processo de *Tradução embasada em Corpora* (TEsC).

Fase B – Elaboração dos corpora na LC

A Fase B (Etapas 1B a 5B. Ver Quadro 1 e Figura 1, acima) de elaboração dos corpora na LC apresentou os mesmos problemas apresentados na Fase A e foram aplicados procedimentos similares para solucioná-los. Na etapa de elaboração da lista de "palavras-chave" na LC (Etapa 1B), foram elaboradas 25 expressões de busca, seus resultados foram armazenados em 25 arquivos .CSV. Após unificação, essa pesquisa gerou um total de 483 links. A Tabela 3 apresenta a síntese dos resultados finais do processo de elaboração dos corpora na LC. Constata-se que a taxa de arquivos eliminados é inferior à do corpus em LP (pt-BR), devido à geração menor de links duplicados (17% < 25%) e a menos arquivos .PDF protegidos com senha (4% < 10%).

Descrição atividade	Nº item	Tipo	% rel.	% abs.	Observações
Coleta SEO "unif.csv"	483	links			
Remover duplicatas	402	links	-17%	-17%	81 links duplicados
Download concluídos	392	.pdf	-2%	-19%	9 downloads impossíveis; 1 arquivos HTML
Conteúdo duplicado	392	.pdf	0%	-19%	0 encontrados
Edit PDF Metadata	375	.pdf	-4%	-22%	11 arquivos protegidos; 6 arquivos inválidos
Some PDF to TXT	373	.txt	-1%	-23%	2 arquivos inválidos
Nº de textos válidos	356	.txt	-5%	-26%	17 arquivos vazios (sem OCR)

Tabela 4 – Síntese final do processo de elaboração dos corpora na LC

Fonte: elaborado pelo autor

A análise do corpus 02_CORPUS_ENO_FR_CUSTOMIZADO (sem limpeza), por meio da ferramenta "Word List" do concordanciador *AntConc*, apresentou os resultados seguintes: 164.837 *Word Types* e 5.140.233 *Word Tokens*. Percebe-se que a composição desse corpus na LC é bastante similar à do corpus na LP, no que se refere à razão dos *Word Tokens* dos corpora na LP e na LC. Porém, a razão dos *Word Tokens* apresenta, na

Tabela 5, maior diferença (24%), o que significaria que o corpus na LC é mais "rico" do ponto de vista lexicológico. Essa observação não pode ser considerada infalível, neste estágio, pois os corpora não foram submetidos a um processo de limpeza avançado, devido às finalidades desse estudo.

Tabela 5 – Comparação estatística dos corpora customizados na LP e na LC
Fonte: elaborado pelo autor

	Corpus LP: pt-BR	Corpus LC: fr-FR	Razão LP/LC
Word Types (Ty)	125885	164837	24%
Word Tokens (To)	5407726	5140233	-5%
Razão Ty/To	2,33%	3,21%	

O corpus elaborado na língua de chegada (LC) está pronto para ser explorado com a concordanciador *AntConc*, no processo de *Tradução embasada em Corpora* (TEsC).

5. Conclusões e encaminhamentos

O princípio de associação das configurações avançadas dos mecanismos de busca (*Google Search Engine*) e das funcionalidades da SEO (*SEOquake*) permitiu a elaboração rápida de dois corpora customizados, um na língua de partida e outro na língua de chegada, resultando em uma média de 130.000 *word types* e de 5.000.000 de *word tokens*. De fato, o trabalho preparatório da elaboração das duas listas de "palavras-chaves" (KW) revelou-se imprescindível para reforçar a aderência do conteúdo dos textos selecionados com a temática geral do texto de partida a ser traduzido. A escolha das ferramentas (gratuitas) usadas contribuiu positivamente para reduzir o tempo gasto durante as tarefas repetitivas e fastidiosas de coletas e de conversões necessárias, por meio do trabalho em lotes (batch). Ademais, esses sucessivos processos de conversões auxiliaram para remover os textos "inválidos".

Os corpora são imediatamente disponíveis e consultáveis pelo tradutor por meio do concordanciador *AntConc* na LP e LC, reduzindo, dessa forma, a frequência de acessos "espontâneos" aos mecanismos de busca, sem nenhuma elaboração clara de uma expressão de busca. Essa metodologia aumenta a concentração do tradutor durante o processo tradutório, proporcionando um acréscimo qualitativo do produto e uma diminuição do tempo de execução da tarefa (futura pesquisa). Ao contrário do que afirma Zanetti (2002), os corpora "pronto para usar" não são necessariamente descartáveis, pois são compostos de um conjunto de textos de qualidade. Eles podem se tornar, com as devidas preparações, corpora de estudo ou de referência, fruto da valorização da atitude proativa do tradutor.

A aplicação da "Análise Multidimensional" (BIBER; PINTO; SARDINHA, 2014) para definir as "dimensões" e os "registros" dos corpora pode-se um encaminhamento de pesquisa promissor para a *Tradução embasada em Corpora* (TEsC). Com efeito, um processo de coleta automatizada, associada à "otimização dos mecanismos de busca" (SEO), bem elaborado, seleciona textos relevantes. A "Análise Multidimensional" pode contribuir para aprimorar, com maior segurança, a aderência do conteúdo dos textos selecionados com a temática geral dos textos de partida e reforçar a eficácia dos corpora em uso.

Referências

BAKER, Mona. *et al. Text and technology in honour of John Sinclair*. Philadelphia: J. Benjamins Pub. Co., 1993.

BERNARDINI, Silvia; BARONI, Marco. *BootCaT:* Bootstrapping corpora and terms from the web. Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), 26-28 May 2004, Lisbon, Portugal, 1313-1316.

BIBER, Douglas; PINTO, Marcia Veirano; SARDINHA, Tony Berber (Org.). *Multi-dimensional analysis*, 25 years on: a tribute to Douglas Biber. Studies in Corpus Linguistics (SCL), volume 60. Amsterdam ; Philadelphia: John Benjamins Publishing Company, 2014.

CALIARI, Vinícius; ROSIER, Jean Pierre; BORDIGNON-LUIZ, Marilde. VINHOS ESPUMANTES: MÉTODOS DE ELABORAÇÃO. ***Evidência - Ciência e Biotecnologia***, [S.l.], v. 13, n. 1, p. 65-77, mar. 2013.

CIMERMANOVÁ, Ivana. Corpus vs dictionary in EFL classes. In: KAČMÁROVÁ, Alena (Org.). *English Matters III*. Prešov (República Eslovaca): Editora da Universidade de Prešov, 2012, p. 65-73.

ENGE, Eric. *et al. A Arte de SEO:* Dominando a Otimização dos mecanismos de busca. Tradução de Tibério Novais e Rafael Zanolli. São Paulo: Novatec, 2012.

FERNEDA, Edberto. Introdução aos Modelos Computacionais de Recuperação de Informação. Rio de Janeiro: Editora Ciência Moderna, 2012.

LAVIOSA, Sara. *Corpus-based translation studies:* theory, findings, applications. Amsterdam; New York, NY: Rodopi, 2002.

OLOHAN, Maeve. *Introducing corpora in translation studies.* London; New York: Routledge, 2004.

PERROTTI-GARCIA, Ana Julia. *O Uso de Corpus Customizado como Fonte de Pesquisa para Tradutores.* Confluências Revista de Tradução Científica e Técnica, Lisboa, v. 3, p. 62-79, 2005.

SARDINHA, Tony Berber. *Linguística de Corpus:* histórico e problemática. DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada, Pontifícia Universidade Católica de São Paulo - PUC-SP, v.16, n.2, p.323-367, 2000.

_____. *Corpora eletrônicos na pesquisa em tradução.* Cadernos de Tradução, Florianópolis, v. 1, n. 9, p. 15-59, jan. 2002.

VIANA, Vander; TAGNIN, Stella (Org.). *Corpora no ensino de línguas estrangeiras*. São Paulo: Hub Editorial, 2011.

_____. *Corpora na tradução.* São Paulo: HUB Editorial, 2015.

ZANETTIN, Federico. DIY Corpora: The WWW and the Translator. In: MAIA, Belinda; HALLER, Jonathan; URLRYCH, Margherita (Org.). *Training the Language Services Provider for the New Millennium*, Porto: Faculdade de Letras, Universidade do Porto, 2002, p. 239-248.

_____. *Translation-Driven Corpora Corpus:* Resources for Descriptive and Applied Translation Studies. Hoboken: Taylor and Francis, 2012.