PROGRAMAÇÃO DE MACROS EM VBA PARA PESQUISAS EM ESTUDOS DA TRADUÇÃO BASEADOS EM CORPUS: DESENVOLVENDO UMA PLANILHA PARA ALINHAMENTO DE CORPUS BILINGUE

Dra. Lilian Fleuri
Dra. Maria Lúcia Vasconcellos

RESUMO: Assim como a Linguística de Corpus, os ETBC estão pautados na Metodoliga de Corpus. Neste âmbito, a ferramenta de análise lexical torna-se também o objeto de estudo (cf. Kennedy, 1998). Partindo de Mason (2008), Garretson (2008), Scott (2008) e Meyer (2002), este artigo explora as possibilidades de programação em Macros VBA para lidar com pesquisas de corpus voltadas aos estudos da tradução. Realiza isso desenvolvendo um Template em um Aplicativo de Planilha, como o MS Excel, para alinhar corpora paralelos bilíngues. Baseando-se nas funções existentes do WordSmith Tools (Scott, 2010), esta planilha é desenvolvida a partir da manipulação de fórmulas e de programação de Macros em VBA. O resultado é uma ferramenta de alinhamento econômica, flexível e adaptável aos objetivos de cada pesquisador. Constata-se que a programação é uma alternativa possível e econômica de manipular corpus, corroborando com Meyer (2002) que aponta que os linguistas serão capazes de desenhar programas específicos de análise de textos, voltados as suas próprias pesquisas.
PALAVRAS-CHAVE: Estudos da Tradução Baseados em Corpus; Corpus paralelo bilíngue; alinhamento em paralelo; programação em Macros VBA.

RESUMEN: Al igual que la Lingüística de Corpus, los ETBC están fundamentados en la Metodología de Corpus. En este ámbito la herramienta de análisis lexical se vuelve también objeto de estudio (cf. Kennedy 1998). A partir de Mason (2008), Garretson (2008), Scott (2008) y Meyer (2002), este artículo explora las posibilidades de programación en Macros VBA para lidiar con investigaciones de corpus direccionadas hacia los estudios de traducción. Eso se realiza con el

desarrollo de un Template en una Aplicación de Plantilla, como el MS Excel, para alinear corpora paralelos bilíngües. Basándose en las funciones existentes en WordSmith Tools (Scott, 2010), esta plantilla es desarrollada a partir de la manipulación de fórmulas de programación de Macros en VBA. El resultado consiste en una herramienta de alineación económica, flexible y adaptable a los objetivos de cada investigador. Se concluye que la programación es una alternativa posible y económica para manipular corpus, corroborando con Meyer (2002), que apunta que los lingüistas serán capaces de diseñar programas específicos de análisis de textos, direccionados hacia su propia investigación.
PALABRAS CLAVE: estudios de traducción fundamentadoss en corpus; corpus paralelo bilíngüe; alineación en paralelo; programación en macros VBA.

O uso do computador na linguística, segundo Mason (2008, p.142), "mudou de simplesmente ser uma glorificada máquina de digitar artigos a ser uma essencial ferramenta de pesquisa". Esta ferramenta passou a ser associada à Metodologia de Corpus (MC) da Linguística de Corpus (LC) e dos Estudos da Tradução Baseados em Corpus (ETBC), sendo aplicada em investigações de descrição e análise linguística em corpus, como as pesquisas desenvolvidas no projeto CORDIALL (Corpus Discursivo para Análises Linguísticas e Literárias)[47]; em elaboração,

[47] O CORDIALL é um corpus multilíngue constituído de textos eletrônicos "selecionados através de critérios específicos relacionados a subprojetos de pesquisa implementados pelos pesquisadores do NET e da UFSC" (CORDIALL, 2010). Atualmente este corpus ultrapassa um milhão de palavras e abarca quatro subcorpora que vêm sendo continuamente atualizados por pesquisas desenvolvidas nas universidades às que está vinculado. Estes subcorpora são (i) corpus paralelo multilíngue; (ii) corpus comparável monolíngue em português brasileiro; (iii) corpus especializado de textos acadêmicos e jornalísticos e (iv) Corpus Processual para Análises Tradutórias, o CORPRAT (Mauri, 2009, p.18). O CORDIALL "fomenta a realização de trabalhos cujo foco principal é a representação de personagens e a apresentação do discurso em textos ficcionais, ambos os aspectos de interesse para a presente pesquisa, que visa a estabelecer uma correlação entre os elementos linguísticos investigados na representação de personagens e na apresentação do discurso com o ponto de vista narrativo" (idem, ibidem). Este projeto é ligado à Universidade Federal de Minas Gerais (UFMG), mais especificamente ao NET (Núcleo de Estudos da Tradução) da FALE (Faculdade de Letras da Universidade Federal de Minas Gerais), e à Universidade Federal de Santa Catarina (UFSC)—PPGET (Programa de Pós-Graduação em Estudos da Tradução) e ao PPGI (Programa de Pós-Graduação em Inglês. Além disso, este projeto é vinculado internacionalmente à Universidad Autonoma de Barcelona e à Universidade de Lisboa. As pesquisas desenvolvidas pelo CORDIALL associam, dentro da disciplina dos ET, duas áreas linguísticas: o arcabouço metodológico da LC e teórico da LSF para descrever a analisar textos em relação tradutória.

descrição e análise de funcionamento de ferramentas tecnológicas computacionais para trabalho com textos, como realiza Scott (2008), Garretson (2008) e Mason (2008); e em outros estudos (cf. Kennedy, 1998, p.9). Este artigo é fruto da proposta defendida na tese de doutorado intitulada "Uma Proposta Metodológica para Compilação de Corpus Paralelo Bilíngue e de Pequena Dimensão" (Fleuri, 2013) que visa expor um modo alternativo de se lidar com as ferramentas computacionais para desenvolver pesquisas em descrições e análises textuais no campo dos Estudos da Tradução Baseados em Corpus. Propõe-se utilizar Suítes de Aplicativos para Escritório, como o MS Office ou o OpenOffice, para programar em Macros VBA tarefas de alinhamento de corpora bilingues e paralelos. Os artigos de Mason (2008), Garretson (2008) e Scott (2008) mostram como a programação vem sendo associada à MC em estudos linguísticos. Como exemplo de associação da programação à pesquisa de corpus, demonstra-se neste artigo a elaboração de um Template em um Aplicativo de Planilhas para realizar a tarefa de alinhamento de corpus paralelo bilíngue. O objetivo deste artigo é exibir possíveis alternativas de se manipular corpora bilíngues sem a necessidade de se ter que adquirir programas comerciais de alinhamento em paralelo, como é o caso do programa WordSmith Tools, amplamente utilizado por investigadores do projeto CORDIALL.

Se computador é o meio pelo qual uma pesquisa de corpus é realizada, ou seja, seu uso é intrínseco à metodologia (Kennedy, 2008, p.5), e se os ETBC são de cunho essencialmente metodológicos (Olohan, 2004, p.3), o computador e as ferramentas de manuseio de corpus são, então, peças-chaves nos estudos de corpus. A evolução dessa tecnologia nos estudos de corpus abriu e continua abrindo espaço para que pesquisas tecnológicas sejam associadas a este estudo, no que se refere à relação das análises linguísticas, de resultados e de métodos e ferramentas aplicados, ou de desenvolvimento de programas de processamento de corpus e de dados. Nota-se que de 1960 a 1990, década em que Kennedy (1998) publica seu livro, *An introduction to Corpus Linguistics*, as tecnologias do computador se aperfeiçoaram enormemente, em termos de armazenamento, velocidade, aplicativos, ferramentas e programas. De 1990 a 2010 as tecnologias se desenvolveram ainda mais. Em duas décadas não apenas evoluiram muito as tecnologias dos PCs e seus sistemas operativos, como também surgiram os dispositivos em formato de tabletes, como Ipad e Iphones (Apple Inc.) e similares, e seus aplicativos. Além disso, foram criados novos programas de processamento

de corpus e de dados que respondem às necessidades das novas tendências de pesquisa e de aplicações às pesquisas. O programa de análise lexical *WordSmith Tools* cf. (Scott, 2010), por exemplo, foi desenvolvido na década de 90 e continua até hoje sendo amplamente utilizado em pesquisas que envolvem metodologia de corpus[48]. Nos anos 2000, Laurence Anthony[49] desenvolveu um programa de concordância *freeware* (gratuito e disponível para baixar online), com funções semelhantes do *WordSmith Tools*, chamado *AntConc*[50] e ainda lançou recentemente, em 2012 e 2013, um programa de concordância paralela (*AntPConc*), de *profiling word* (*AntWordProflier*), de análise estrutural de text (*AntMover*), entre outros. Ainda na entrada dos anos 2000, Barlow (2003) lançou um programa de concordância paralela chamado *ParaConc*, também amplamente utilizado em pesquisas que realizam alinhamento de corpus e corcordância em corpus alinhados.

Dentro do viés metodológico, observar-se que pesquisas que envolvem metodologia de corpus não realizam estudos apenas em análise textuais e linguísticas, elas também preocupam-se com a avaliação e desenvolvimento de métodos e de ferramentas computacionais, conforme explica Kennedy (1998, p.9) "o trabalho na linguistica de corpus é atualmente associados a diversas atividades", divididas em quatro grupos de pesquisadores:

> O primeiro grupo de pesquisadores envolve os elaboradores de corpus ou compiladores de corpus. Tais acadêmicos se dedicam a desenhar e compilar corpora, a coleção de textos e sua preparação e armazenamento para análises posteriores. / **O segundo grupo de pesquisadores se dedicam a desenvolver ferramentas para analisar corpora.** Importantes contribuições ao desenvolvimento de software, especialmente para a análise sintática de corpora, têm sido associadas particularmente, mas não exclusivamente, a pesquisadores em Linguística Computacional. Esses pesquisadores têm se dedicado ao uso de corpora para

[48] O site de Mike Scott aponta algumas pesquisas desenvolvidas com e sobre o WordSmith Tools: (que usa WST) http://www.lexically.net/wordsmith/corpus_linguistics_links/papers_using_wordsmith.htm (que cita o WST): http://www.lexically.net/publications/citing_wordsmith.htm

[49] http://www.antlab.sci.waseda.ac.jp/software.html (Fonte acessada em julho/2013)

[50] Confira http://research.ncl.ac.uk/decte/toon/assets/docs/AntConc_Guide.pdf para obter um guia explicativo de utilização e descrição do AntConc.

> desenvolver, entre outras coisas, algoritmos para processamento de linguagem natural e o modelamente de teorias linguísticas. / O terceiro grupo de pesquisadores envolve linguistas descritivos cuja principal preocupação é fazer uso de corpora computadorizados para descrever confiavelmente o lexicon e a gramática da língua, (...). É o aspecto probabilistico de corpus baseado em estudos descritivos da língua que especialmente distingue-os de convencionais campos de estudos descritivos em linguística ou lexicografia. (...). / A quarta área de atividade, que está entre os resultados mais inovativos da revolução de corpus, é a exploração de descrição linguística baseada em corpus no uso de uma variedade de aplicações, tais como aprendizagem e ensino de línguas e processamento de linguagem natural por máquinas, incluindo o reconhecimento da fala e a tradução. (Kennedy, 1998, p.9, grifo meu)

Esse enunciado explica que atualmente a maioria dos pesquisadores concentram-se na descrição linguística baseada em corpus, embora haja grupos que se preocupam em estudar desenho de corpus, e outros que se concentram em estudar métodos para análise e processamento de texto. Olohan (2004) também observa que o caráter metodológico de pesquisas em tradução baseada em corpus permite, por exemplo, imaginar pesquisas sobre a "aplicação da metodologia a diferentes tipos de tradução". Atualmente observa-se uma escassez de estudos na área dos estudos de corpus que se preocupem em analisar e propôr métodos e ferramentas (cf. Mason, 2008). Olohan (2004, p.16) aponta que a orientação não-prescritiva das estruturas teóricas para métodos de corpus em estudos da tradução pressupõem o desenvolvimento de pesquisas majoritariamente enfocadas em análises linguísticas de traduções. Tal lacuna é também observada e constatada na tese de Fleuri (2013).

Entre os escassos artigos que exploram o enforque metodológico das ferramentas computacionais aos estudos de corpus encontram-se os artigos de Mason (2008), Garretson (2008) e Scott (2008), publicados em 2008 na revista *International Journal of English Studies*, no volume especialmente dedicado a questões relativas ao uso e desenvolvimento de softwares de análise linguísticas, intitulado *Monograph: Software-aided*

Analysis of Language. Esses três artigos discutem programação por linguistas para desenvonver tarefas voltadas às suas pesquisas. O presente artigo dialoga com Oliver Mason (2008), *Developing Software for Corpus Research*, quanto à proposta de desenvolver um método economicamente acessível de processamento de corpus, sem necessidade de ser um programador. Quanto ao artigo de Mike Scott (2008), *Developing WordSmith*, o diálogo se estabelece entre linguistas de corpus com interesse e iniciativa de pensar na metodologia de corpus a partir das ferramentas de análise. Como linguista ele começou a programar, a princípio de modo rudimentar, o programa que hoje é um dos mais usados nos estudos de corpus—o *WordSmith Tools*. Finalmente, no artigo intitulado *Desiderata for Linguist Software Design*, Gregory Garretson (2008) apresenta uma série de diretrizes para pesquisadores a procura de software de análise linguística e para programadores que desenham tais software. Esse artigo se comunica com a presente pesquisa quando estabelece um diálogo entre pesquisadores e programadores e a colaboração que um oferece ao trabalho do outro—relacionando principalmente o papel do linguista de corpus na tarefa de programação. Também apresenta uma reflexão sobre os elementos a se considerar, durante a etapa de planejamento da pesquisa, na adoção de um software ou na criação de um software específico para a pesquisa.

Mason (2008) aponta que estudos baseados em corpus "dependem muito da programação de computador e infelizmente poucos pesquisadores são capazes de desenhar seu próprio software, tendo então que depender de programas existentes para fazer seus processamentos" (p.155). Os programas existentes apresentam em geral algumas limitações (idem, p.141), não respondendo às necessidades da pesquisa. Neste caso, levam as pesquisas a se adaptarem aos programas existentes, ou levam o pesquisador a se apoiar em um segundo software ou a realizar grande parte do trabalho manualmente. Nesse sentido, conhecer as capacidades dos software disponíveis e saber escolher qual programa e ferramentas usar é essencial no planejamento da pesquisa em corpus. As escolhas, ressalta Mason, "dependem da natureza da análise e da disponibilidade do programa" (idem, p.142). Saber que dentre essas escolhas há também a possibilidade de se criar programas que respondam às necessidades da sua pesquisa é uma importante informação que pode ajudar muito o pesquisador a planejar seu estudo.

Nessa linha de raciocínio e para fins de atender às necessidades de cada pesquisador, é possível afirmar que nem toda tarefa de programar

precisa ser complexa e envolver um projeto com profissionais de áreas de programação. Mason argumenta que "é facilmente possível adquirir algumas habilidades necessárias para se criar pequenos programas em linguagem de *script*, acelerando muito o processamento de texto e de dados" (idem, p.155). Ele exemplifica essa possibilidade, explicando e demonstrando três possíveis linhas de comando escritas no 'terminal' do Sistema Operativo do Windows para: criar lista de frequência, contar número de palavras terminadas em *–ing*, por exemplo, e encontrar palavras em comum entre dois textos (idem, p. 150-153). Para realizar essas tarefas, como argumenta Mason (idem, ibdem), pode-se usar uma linguagem simples, acessível a iniciantes em processamento, lidando com o sistema operativo do Windows (MS-DOS), familiar a maioria os pesquisadores que utilizam o computador como editor de textos. O que acontece geralmente é que grande parte dos pesquisadores não têm consciência das opções disponíveis e não sabem por onde começar a buscar informação para desenvolver uma simples programação. O artigo de Mason oferece aos pesquisadores em corpus um ponto de partida para o planejamento metodológico de suas pesquisa, mostrando uma forma de criar, de modo econômico, flexível e personalizado um programa que responda as suas necessidades, ao apresentar algumas possibilidades de trabalho dentro de Sistemas Operativo de Disco. De fato, menciona Mason, a maioria dos softwares de processamento de corpus disponíveis no mercado foram criados por "entusiastas amadores" (idem, p.144).

Ao contrário do que argumenta Mason (2008), Garretson (2008) sustenta que programar é uma atividade conjunta entre programador e linguistas de corpus. Isso ocorre porque Garretson vê a tarefa de programar como algo realizado por cientistas da computação e programadores profissionais, que lidam com sistemas operativos diferentes aos mencionados por Mason (2008). Apesar de sustentar seu artigo na premissa de que o desenvolvimento de um software deve ser um trabalho conjunto entre linguista e programador, Garretson aponta como uma situação ideal o fato de o próprio linguista ser programador de suas próprias ferramentas, pois afirma que "acredito que linguistas deveriam ser encorajados a aprender técnicas de programação (cf. Biber et al. 1998, p.254-256) e posso testemunhar que é possível que software linguísticos sejam desenvolvidos por indivíduos que não tenham um background em ciências da computação" (Garretson, 2008, p.75).

Garretson (2008) apresenta uma desiderata para desenhar software partindo de cinco questões que um pesquisador em estudos de

corpus deve considerar antes de iniciar a pesquisa, em relação a (i) os requerimentos de software; (ii) o formato dos dados; (iii) o uso de software existentes; (iv) alguns problemas de softwares prontos; e (v) a criação de um software específico para a pesquisa. Quanto ao último ítem, deve-se considerar como adquirir a linguagem de programação ou a ajuda de um programador, gratuitamente através de acadêmicos ou estudantes voluntários que saibam programação, através de colaborações de pessoal da área de ciências da computação ou do departamento técnico, ou através de acordos e projetos conjuntos de pesquisas com áreas da linguagem natural ou linguística computacional. Ou então, considerar a contratação de um programador, ou ainda investir que alguém envolvido no projeto comece a aprender a programar; a primeira opção pode ser muito cara e a segunda demorada (idem, p.74-75).

Ele chama de 'desiderata' os objetivos e princípios a se considerar quando se trabalha em um projeto de software linguístico. Evitando-se detalhes e discussões técnicas, estas desideratas estão exibidas em grupos temáticos:

i) Princípios de desenho geral: (a) planejar o tempo em relação ao desenvolvimento de um software, (b) aproveitar a que já existe, tornar o programa flexível, (c) buscar acomodar o programa a uma grande gama de usuários, (d) tornar o programa mais claro que engenhoso;

ii) Teoria linguística e dados linguísticos: separar os *markups* das anotações. Chama-se de *markup* "as informações extras adicionadas durante a criação do texto, que normalmente seriam perdidas" (idem, p.78) e de anotação "o registro de uma análise teórica dos dados" (idem). Aconselha-se inclusive que os *markup* sejam anotados em documentos diferentes das anotações; e permitir que os usuários forneçam suas próprias categorias, não limitando a análise desnecessariamente;

iii) A interface de usuário: (a) deixar o desenho da interface o menos técnica possível, evitando deixá-lo confuso; oferecer um conjunto de opções, evitando que o usuário tenha que saber o que digitar; (b) usar um usuário gráfico, mais do que uma interface de linha de comando; (c) uma interface gráfica não é apenas mais fácil de usar, como também familiar à maioria dos usuários" (idem, p.80); (d) apresentar os resultados em estágios; (e) tornar fácil de instalar e fazer *updates* no programa.

iv) Documentação: (a) documentar assim que se começa a programar; (b) pecar por oferecer informação demais que de menos; (c) documentar de modo que o leitor vá compreender; (d) documentar o código para que outros programadores tenham acesso;

v)Teste, depuração e tratamento de erro: Garretson enfatiza que todo programa tem que ser testado e a maioria dos *bugs* devem ser removidos antes de disponibilizados aos usuários.

vi) Capacidades de busca: equilibrar complexidade (muitas opções de busca) e velocidade na busca;

Conforme o ideal mencionado por Garretson (2008), Mike Scott (2008) é um exemplo de pesquisador que associa seu hobby de programação aos conhecimentos de linguista de corpus. No artigo intitulado *Developing WordSmith*, descreve sua trajetória, como linguista aplicado ao ensino de línguas e a corpus, na programação do seu software de análise lexical de textos, *WordSmith Tools*. Ele explica, em uma linguagem acessível a linguistas de corpus o contexto em que se deu o desenvolvimento do programa e os principios em quais se baseia seu desenho. Comenta que o que o impulsionou a desenvolver tal programa foi o desejo de praticar um hobby auto-didata. Ao elaborar ferramentas ele poderia entender as quais poderia relacionar, suas próprias teorias, trabalho e interesses (idem, p.96). Através deste relato, Scott mostra que qualquer linguista de corpus pode escrever *scripts* para se analisar textos (idem, p.104), concordando com Meyer (2002, p.141) que diz:

> Muito provavelmente a próxima geração de linguista de corpus vai ter um *background* de programação muito melhor. Assim, estes linguistas de corpus serão capazes de aplicar o conhecimento da linguagem *Perl* ou *Visual Basic* para escrever *scripts* específicos para a análise de textos, e a medida que estes *scripts* vão se proliferando, poderão ser passados para frente e talvez tornar obsoleta a necessidade de desenhar programas específicos de análise de textos. (Meyer apud Scott, 2008, p.103).

Entretanto ressalta que poucos linguistas de corpus de fato programam ou analisam métodos em relação à funcionalidade das ferramentas computacionais. A razão disso é explicada por analogia: o

linguista de corpus geralmente se coloca para o computador, como o motorista se coloca para o carro: a maioria deles estão mais interessados na paisagem que no carro em si. Ou seja, o linguista se preocupa em usar a ferramenta computacional para executar a análise, e não tanto em pensar como melhorar a ferramenta para se encaixar à necessidade de sua análise. Como motoristas, "enquanto o carro está funcionando permanecemos na estrada, se queremos subir uma colina, tendemos a ir caminhando" (idem, p.104). Esse comentário corrobora a menção de Olohan (2004), que apesar de os estudos de corpus assumirem os estudos metodológicos de corpus, como a descrição das ferramentas e seu funcionamento, a maioria dos linguistas parecem estar mais interessados em descrever as particularidades do corpus em si que as ferramenta de corpus; mesmo que o funcionamento e manuseio desta influencie diretamente no resultado das descobertas linguísticas.

A LC assim como os ETBC, por ser um campo essencialmente metodológico (Olohan, 2004; Kennedy, 1998), permite que o foco do estudo se direcione ao método e suas ferramentas em si, além da usual aplicação da metodologia na descrição e análise textual. Estudos sobre métodos e ferramentas de análise linguística são ainda escassos na área da LC e dos ETBC, conforme apontou o encontro de linguistas de corpus em Georgetown University em 1993 (Kennedy, 1998). Tal escassez faz com que a programação se torne um estudo mais matemático (Mason, 2008) e voltado ou associado à linguística computacional (Kennedy, 1998). Entretanto, Mason (2008), Scott (2008) e Garretson (2008) mostram que examinar o funcionamento de ferramentas e de suas aplicações aos métodos de corpus não se tem que ser tarefa exclusiva do cientista computacional. De fato, não é necessário que o pesquisador seja um *expert* em linguagem de programação e em desenvolvimento de ferramentas para análise textual para realizar estudos nessa área: qualquer linguista de corpus pode perfeitamente desenvolver esta tarefa, como prova Scott (2008). Segundo Garretson (2008), pensar a ferramenta e a possibilidade de programar ou não deve ser uma das considerações a se levar em conta na hora de se planejar uma pesquisa em corpus. Afinal, não há realmente como evitar a entrada e ou a consolidação dessa atividade de pesquisa nos estudos de corpus, como prevê Meyer (apud Scott, 2008), no futuro todos os linguistas de corpus serão um pouco programadores e seus *scripts* serão passados adiante de modo que um dia possam apoiar suas pesquisas em programas escritos e adaptados pelos próprios pesquisadores. Vale destacar, entretanto, que apesar de defender esta ideia e de dialogar com a linguagem de programação, este artigo

oferece alternativas de programação básica em *VBA* e *Macros*[51] para desenvolver um *Template* em um Aplicativo de Planilhas, flexível e passível de ser reprogramado e adaptado para realizar alinhamento paralelo de corpora paralelo bilíngue.

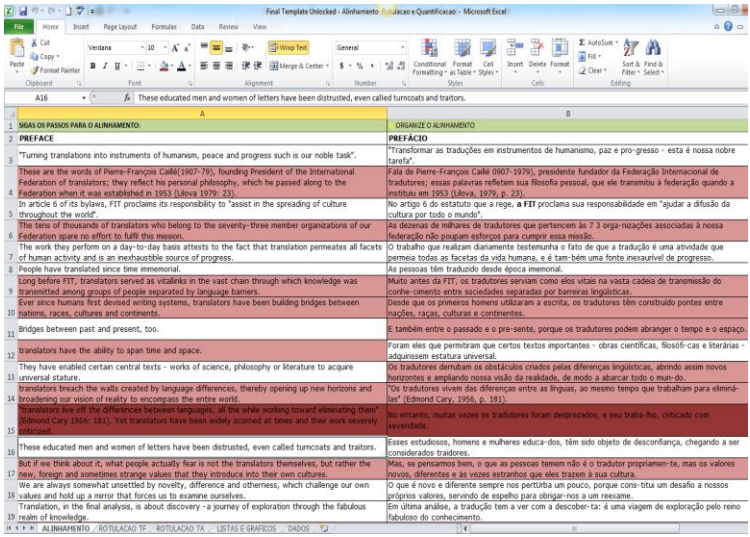
O alinhamento de corpora paralelos é uma prática muito utilizada em pesquisas em ETBC, que pode ser realizado manualmente ou semi-automaticamente com auxílio de programas, como o *WordSmith Tools* (Scott, 2010) e o *ParaConc* (Barlow, 2003). Se por um lado a tarefa de realizar o alinhamento manualmente é lenta e laboriosa, realizá-la semi-automaticamente pode ser bastante cara. Com o intuito de tornar esta tarefa dinâmica e econômica, exploram-se os recursos de software acessíveis aos usuários dos Sistemas Operativos de Disco, como Windows, Macquintosh e Linux. No caso aqui proposto, é utilizado o aplicativo de planilhas do MS Office, o MS Excel[52], sendo o recurso de programação explorado o de elaboração de Fórmulas associada a programação de Macros em VBA.

Tendo em mente algumas das desideratas de Garretson (2008) e inspirando-me na proposta de Mason (2008), proponho neste artigo e em minha tese de doutorado um Template de Alinhamento de corpus paralelo bilíngue (inglês/português). Este Template[53] executa as funções de 'juntar' e 'quebrar' sentenças e de 'buscar por pares desalinhados', funções também executadas pelo *Viewer and Aligner* do *WordSmith Tools*. No Aplicativo de Planilhas, o corpus é alinhado em sentenças ou parágrafos lado-a-lado, sendo texto na L1 posicionado na coluna A da planilha e o texto na L2, na coluna B, conforme ilustrado na Figura 1:

[51] A definição de Macros, dada pelo site oficial da Microsoft Office é: "Um Macro é a coletânea de comandos que se pode aplicar com um único click. Eles podem automatizar quase tudo que se faz no programa que se está utilizando e até mesmo permitir de se fazer coisas que nunca se imaginou ser possível" (Fonte consultada em setembro de 2012: http://office.microsoft.com/en-us/help/).

[52] Ou seu equivalente do OpenOffice, se a ideia é utilizar um software aberto, como o Linux.

[53] Este Template pode ser baixado gratuitamente no site: https://sites.google.com/site/lilianjfleuri/



	A	B
1	SIGAS OS PASSOS PARA O ALINHAMENTO:	ORGANIZE O ALINHAMENTO
2	PREFACE	PREFÁCIO
3	"Turning translations into instruments of humanism, peace and progress such is our noble task".	"Transformar as traduções em instrumentos de humanismo, paz e pro-gresso - esta é nossa nobre tarefa".
4	These are the words of Pierre-François Caillé(1907-79), founding President of the International Federation of translators; they reflect his personal philosophy, which he passed along to the Federation when it was established in 1953 (Lilova 1979: 23).	Falá de Pierre-François Caillé 0907-1979), presidente fundador da Federação Internacional de tradutores; essas palavras refletem sua filosofia pessoal, que ele transmitiu à federação quando a instituiu em 1953 (Lilova, 1979, p. 23).
5	In article 6 of its bylaws, FIT proclaims its responsibility to "assist in the spreading of culture throughout the world".	No artigo 6 do estatuto que a rege, **a FIT** proclama sua responsabilidade em "ajudar a difusão da cultura por todo o mundo".
6	The tens of thousands of translators who belong to the seventy-three member organizations of our Federation spare no effort to fulfil this mission.	As dezenas de milhares de tradutores que pertencem às 7 3 orga-nizações associadas à nossa federação não poupam esforços para cumprir essa missão.
7	The work they perform on a day-to-day basis attests to the fact that translation permeates all facets of human activity and is an inexhaustible source of progress.	O trabalho que realizam dianamente testemunha o fato de que a tradução é uma atividade que permeia todas as facetas da vida humana, e é tam-bém uma fonte inexaurível de progresso.
8	People have translated since time immemorial.	As pessoas têm traduzido desde época imemorial.
9	Long before FIT, translators served as vitallinks in the vast chain through which knowledge was transmitted among groups of people separated by language barriers.	Muito antes da FIT, os tradutores serviam como elos vitais na vasta cadeia de transmissão do conhe-cimento entre sociedades separadas por barreiras lingüísticas.
10	Ever since humans first devised writing systems, translators have been building bridges between nations, races, cultures and continents.	Desde que os primeiros homens utilizaram a escrita, os tradutores têm construído pontes entre nações, raças, culturas e continentes.
11	Bridges between past and present, too.	E também entre o passado e o pre-sente, porque os tradutores podem abranger o tempo e o espaço.
12	translators have the ability to span time and space.	Foram eles que permitiram que certos textos importantes - obras científicas, filosófi-cas e literárias - adquirissem estatura universal.
13	They have enabled certain central texts - works of science, philosophy or literature to acquire universal stature.	Os tradutores derrubam os obstáculos criados pelas diferenças lingüísticas, abrindo assim novos horizontes e ampliando nossa visão da realidade, de modo a abarcar todo o mun-do.
14	translators breach the walls created by language differences, thereby opening up new horizons and broadening our vision of reality to encompass the entire world.	"Os tradutores vivem das diferenças entre as línguas, ao mesmo tempo que trabalham para elimná-las" (Edmond Cary, 1956, p. 181).
15	Translators live off the differences between languages, all the while working toward eliminating them" (Edmond Cary 1956: 181). Yet translators have been widely scorned at times and their work severely criticized.	No entanto, muitas vezes os tradutores foram desprezados, e seu traba-lho, criticado com severidade.
16	These educated men and women of letters have been distrusted, even called turncoats and traitors.	Esses estudiosos, homens e mulheres educa-dos, têm sido objeto de desconfiança, chegando a ser considerados traidores.
17	But if we think about it, what people actually fear is not the translators themselves, but rather the new, foreign and sometimes strange values that they introduce into their own cultures.	Mas, se pensarmos bem, o que as pessoas temem não é o tradutor propriamen-te, mas os valores novos, diferentes e às vezes estranhos que eles trazem à sua cultura.
18	We are always somewhat unsettled by novelty, difference and otherness, which challenge our own values and hold up a mirror that forces us to examine ourselves.	O que é novo e diferente sempre nos pertúrba um pouco, porque cons-titui um desafio a nossos próprios valores, servindo de espelho para obrigar-nos a um reexame.
19	Translation, in the final analysis, is about discovery -a journey of exploration through the fabulous realm of knowledge.	Em última análise, a tradução tem a ver com a descober-ta: é uma viagem de exploração pelo reino fabuloso do conhecimento.

Cada uma das três funções desenvolvidas demandou uma programação de Fórmulas associadas aos Macros em VBA[54]. A função de "juntar", gravada no shortcut Ctrl+J, leva o texto da célula de baixo (ex. A2) ao final do texto da célula de cima (ex. A1), sem alterar a ordem das linhas da coluna ao lado. Esta função é realizada pelo seguinte Macro:

```
Sub JoinCellsMoveup()
    If ActiveCell.Row > 1 Then
        ActiveCell.Offset(-1, 0).Value = ActiveCell.Offset(-1, 0).Text
& " " & ActiveCell.Text
        ActiveCell.Delete
    End If
End Sub
```

A função de "quebrar" sentença divide o texto de uma célula em duas células, sem alterar a ordem das linhas da coluna ao lado. Para se realizar esta função deve-se levar o cursor ao ponto do texto que se deseja quebrar; acionar o comando Alt+Enter (shortcut do MS Excel) e em seguida o comando Ctrl+S (shortcut gravado). Esta função é realizada pelo seguinte Macro:

```
Sub SplitCell()
```

[54] Se nesta fase de estudo são consultados fóruns. O fórum mais acessado foi http://answers.microsoft.com/en-us ao qual o usuário tem que se registrar antes de realizar suas perguntas.

```
    Dim s As String, v As Variant, l As Long
        s = ActiveCell.Value
        v = Split(s, vbLf)
        l = UBound(v) - LBound(v) + 1
    If l > 1 Then
        ActiveCell.Offset(1, 0).Resize(l - 1, 1).Insert
Shift:=xlShiftDown
        ActiveCell.Resize(l, 1).Value = Application.Transpose(v)
    End If
End Sub
```

A função de "buscar por pares desalinhadas", que automatiza o processo de busca por sentenças desalinhadas, foi gerada a partir da criação prévia de uma fórmula. A criação da fórmula contou com a constatação de que em um corpus bilíngue português e inglês as sentenças que formam uma unidade desalinhada apresentam uma diferença absoluta de dez palavras. A partir desta constatação, elabora-se uma fórmula para contar o número de palavras em uma célula e outra para buscar pares de linhas A/B que apresentam uma diferença de 9 palavras absolutas (-9 ou +9)[55] entre as células A e B. As fórmulas são as seguintes:

Equação 1: Fórmulas para se buscar desalinhamento entre textos em Português e Inglês — Pasta Alinhamento no *Template MS Excel*

Função	Fórmula
Contar palavras de uma determinada célula da coluna A.	=LEN(A1)-LEN(SUBSTITUTE(A1," ",""))+1
Buscar pares de linhas A-B que apresentem uma diferença de 9 palavras absolutas (-/+9)	=IF(ABS((LEN(A1)-LEN(SUBSTITUTE(A1," ",""))+1)-(LEN(B1)-LEN(SUBSTITUTE(B1," ",""))+1))>=9,"MISMATCH",0)

A partir desta constatação, programa-se o seguinte Macro em VBA, gravado com o *shortcut* Ctrl+M:

```
Sub FindMismatch()
    Dim r As Long
    Dim m As Long
    m = Range("A:B").Find(What:="*", SearchOrder:=xlByRows, _
```

[55] Diminui-se uma palavra para garantir que todos os pares desalinhados fossem encontrados.

```
        SearchDirection:=xlPrevious).Row
    For r = ActiveCell.Row + 1 To m
        If Abs(UBound(Split(Range("A" & r).Value)) - _
            UBound(Split(Range("B" & r).Value))) >= 9 Then
            Application.Goto Range("A" & r), True
            Exit For
        End If
    Next r
End Sub
```

O processo de alinhamento é concluído quando não há mais ocorrências de desalinhamentos. O pesquisador pode optar em formatar o *layout* do alinhamento, por exemplo, destacando com cor diferente as linhas em branco, para melhor identificar as ausências de tradução ou de texto-original no alinhamento. Esses processos de formatação em geral encontram-se disponíveis nos comandos do próprio MS *Excel*. O resultado final é dois textos alinhados lado-a-lado.

Em pesquisas em ETBC, relatam-se o uso dos aplicativos de planilhas, como o MS *Excel* e o *OpenOffice.Calc.org*, para desempenhar funções básicas de organização de textos em tabelas e/ou de gráficos (cf. Zanella, 2006), a contabilização de dados coletados (cf. Fleuri, 2006), para organização e revisão de rótulos (cf. Feitosa, 2005), o alinhamento manual corpus bilíngue (cf. Alves, 2007) entre outras. A tese "Uma proposta metodológica para compilação de corpus paralelo bilíngue de pequena dimensão" (Fleuri, 2013) demonstrou que os aplicativos de planilhas podem também ser útil no trabalho com corpora paralelos bilíngues de pequena dimensão, para alinhar textos, por exemplo, e que a associação de programação para manipular corpus em ETBC pode significar a emancipação de programas de análise lexical pagos assim como a flexibilização da manipulação do corpus durante a análise lexical, de acordo com os objetivos da pesquisa

Muitas pesquisas que utilizam o *WordSmithTool* (WST) (i.e. Fleuri, 2006; Pires, 2009; Fernandes, 2009; Morinaka, 2005; Zuniga, 2006) relatam a praticidade de se realizar concordâncias automáticas e de obterem dados estatísticos do corpus em segundos. Entretanto, também relatam dificuldades em alinhar textos bilíngues de modo completamente automático e necessidades de se passar o alinhamento gerado no *WST* a outro documento (em geral o *MS Word*). Ambos os tipos de programas (de análise lexical e os aplicativos de planilhas) demandam um tempo de

aprendizagem. O *WST* requer um estudo e uma prática de suas funções, enquanto um Aplicativo de Planilha, como o MS Excel, demanda estudo de programação de *Macros* e de fórmulas para a elaboração de um novo *Template* ou a adaptação deste oferecido. Possivelmente o caráter autoexplicativo dos programas de concordância lexical e o fato de estarem inseridos no campo da Linguística de Corpus, faz com que muitos linguistas e tradutores escolham lidar com métodos de manipulação de corpus no *WordSmithTools* e escolham Aplicativos de Planilhas para organizar, de modo manual, textos (ou fragmentos dos corpora) e dados em gráficos e tabelas para cumprir determinados propósitos de sua pesquisa.

De acordo com Kennedy (1998), o método de pesquisa aplicado na LC é associado aos processos de compilar e elaborar corpus; descrever elementos linguísticos/textuais através de corpora computadorizados; realizar descrição linguística em corpus para aplicar à aprendizagem e ensino de línguas e ao processamento de linguagem natural por máquinas; e desenvolver ferramentas para analisar corpora. Essa última atividade, em que se enquadra a presente pesquisa, é descrita por Mason (2008), Scott (2008) e Garretson (2008). A linguística de corpus assim como os ETBC por serem um campo essencialmente metodológico (Olohan, 2004; Kennedy, 1998), permite que o enfoque do estudo se direcione ao método, além da usual aplicação da metodologia na descrição e análise textual em si. Estudos sobre métodos e ferramentas de análise linguística são ainda escassos na área da LC e dos ETBC, conforme apontou o encontro de linguistas de corpus em Georgetown University em 1993 (Kennedy, 1998). Tal escassez faz com que a programação de software se torne um estudo mais matemático (Mason, 2008) e voltado ou associado à linguística computacional (Kennedy, 1998). Entretanto, Mason (2008), Scott (2008) e Garretson (2008) mostram que examinar o funcionamento de ferramentas computacionais e de suas aplicações aos métodos de corpus não tem que ser tarefa exclusiva do cientista computacional. De fato, não é necessário que o pesquisador seja um especialista em linguagem de programação e em desenvolvimento de ferramentas para análise textual para realizar estudos nessa área: qualquer linguista de corpus pode perfeitamente desenvolver esta tarefa, como prova Scott (2008). Segundo Garretson (2008) e Barnbrook (1996), pensar a ferramenta e a possibilidade de programar ou não deve ser uma das considerações a se levar em conta na hora de se planejar uma pesquisa em corpus. Como prevê Meyer (apud Scott, 2008),

no futuro todos os linguistas de corpus serão um pouco programadores e seus scripts serão passados adiante de modo que um dia possam apoiar suas pesquisas em programas escritos e adaptados pelos próprios pesquisadores. A proposta de se utilizar a programação de Macros em VBA em Aplicativos de Planilhas, como a desenvolvida no *Template* MS *Excel*, visa oferecer um modelo básico de trabalho com corpus em paralelo que possa ser adaptado aos objetivos de cada pesquisa. A intenção da criação deste *Template* é mostrar uma possível alternativa de manipulação de corpus paralelo bilíngue. No caso aqui exposto, demonstrou-se como realizar o alinhamento de um corpus paralelo bilíngue em um Aplicativo de Planilha, de modo gratuito, realizando tarefas semelhantes às presentes no *WordSmith Tools* (Scott, 2010). Assim como esta (FLEURI, 2013), outras funções podem ser programadas, permitindo a criação de novas formas de manuseio de corpus, de modo a atender às necessidades individuais de pesquisas em corpus.

BIBLIOGRAFIA

ALVES, D. **Aspectos da Representação do Discurso em Textos Traduzidos:** Os Verbos de Elocução Neutros. Belo Horizonte: Faculdade de Letras UFMG, 2006.

BAKER, M. Corpus Linguistics and Translation Studies: Implications and Applications. In: BAKER, M.; FRANCIS, G.; TOGNINI-BONELLI, E. **Texts and Technology:** in Honour of John Sinclair. Amsterdam and Philadelphia: John Benjamins, 1993. p. 233-250.

BAKER, M. **Corpora in Translation Studies:** An Overview ans Some Suggestions for Future Research. Amsterdam: John Benjamins B.V., 1995.

BARLOW. Paraconc.pdf. **ParaConc:** A Concorder for Parallel Texts, 2003. Disponivel em: <http://www.athel.com/paraconc.pdf>.

BARNBROOK, G. **Language and Computers:** A practical introduction to the computer analysis of language. Edinburgh: Edinburgh University Press, 1996.

BIBER, D. M. Methodological issues regarding corpus-based analyses of linguistic variation. **Literary and Linguistic Computing**, p. 257-269, 1990.

BOWKER, L. Towards a Methodology for Corpus-Based Approach to Translation Evaluation. **Meta: journal des traduteurs/ Meta: Translator's Journal**, v. 46, p. 345-364, 2001.

CORDIALL, P. Letras UFMG, 24 mar. 2010. Disponivel em: <http://www.letras.ufmg.br/net/cordiall/portugues/projeto.htm>.

FEITOSA, M. **Uma Proposta de Anotação de Corpora Paralelos com Base na Linguistica Sistêmico-Funcional**. Belo Horizonte: Programa de Pós-Graduação da Faculdade de Letras/UFMG, 2005.

FERNANDES, A. **Black into white and preto no branco:** can you tell one's colour by the company one keeps? Florianopolis: UFSC/PPGI, v. (Dissertação de Mestrado), 2009.

FERNANDES, L. Corpora in Translation Studies: Revising Baker's Typology. **Fragmentos**, v. 30, p. 87-95, 2006.

FLEURI, L. **Uma proposta metodológica para compilação de corpus paralelo bilíngue de pequena dimensão**. Florianópolis: Universidade Federal de Santa Catarina, v. Orientadora Maria Lúcia Vasconcellos, 2013.

FLEURI, L. J. **O Perfil Ideacional dos Itens Lexicais Translator/Tradutor em "Translator Through History"**. Florianópolis: Programa de Pós-Graduação em Estudos da Tradução/UFSC, 2006.

GARRETSON, G. Desiderata for Linguistic Software Design. **Internatina Jornal of English Studies (IJES)**, v. 8 (1), p. 67-94, 2008.

KENNEDY, G. **An Introduction to Corpus Linguistics**. New York: Logman, 1998.

KRUGER, A. CORPUS-BASED TRANSLATION RESEARCH: Its development and implications for general, literary and Bible translation. **African Journals Online**, p. 70-106, 2002. Disponivel em: <http://www.ajol.info/index.php/actat/article/view/5455>.

LEECH. Corpora and theories of linguistic perfomance. In: _____ **Startvik**. [S.l.]: [s.n.], 1992. p. 105-122.

LEECH, G.; WEISSER, M. Generic Speech Act Annotation for Task-Oriented Dialogue. **Proceedings of the Corpus Linguistics 2003 Conference**, Lancaster, 2003.

MASON, O. Developing Software for Corpus Research. **International Journal of English Studies (IJES)**, v. 8 (1), p. 141-156, 2008. Monograph: Software-aided analysis of Language.

MORINAKA, E. M. **Gabriela, Cravo e Canela and its (re)textualization in English:** Representation Through Lexical Relation. Florianopolis: Pós-Graduação em Letras Inglês/UFSC, 2005.

OLOHAN, M. **Introducing Corpora in Translation Studies**. London/New York: Routledge, 2004.

PIRES, T. **The construal of Bishop's Ideational profile in Flores Raras e Banalíssimas and Rare and Commonplace Flowers:** a corpus-based Translation Study. Florianopolis: UFSC/PPGI, v. (Dissertação de Mestrado), 2009.

SCOTT, M. Developing WordSmith. **International Journal of English Studies (IJES)**, v. 8 (1) , p. 95-106, 2008.

SCOTT, M. WordSmith Tools: software for finding word patterns. **WordSmith Tools:** software for finding word patterns, 2010. Disponivel em: <http://www.lexically.net/downloads/version5/HTML/?getting_started.htm>.

SCOTT, M. WordSmith Tools, 27 junho 2010b. Disponivel em: <http://www.lexically.net/downloads/version6/HTML/?wshell.htm >.

STUBBS, M. **Text and Corpus Analysis**. Oxford: Blackwell Publisher, 1996.

ZANELLA, A. **Mapeamento Macro e Micro estrutural da retextualização de resumos on-line:** Estudo da Transitividade de abstracts biomédicos. Florianopolis: UFSC/PGET, v. (Dissertação de Mestrado), 2006.

ZUNIGA, G. **Construing the Translator in "Becoming a Translator" and "Construindo o Tradutor":** a Case Study Based on Corpus and Systemic Linguistics. Florianopolis: Pós-Graduação em Letras Inglês/UFSC, 2006.