ABORDAGEM HEURÍSTICA DAS LINGUAGENS DE ESPECIALIDADE COM RECURSO À LINGUÍSTICA DE *CORPUS*: CASO DE ESTUDO EM LINGUAGEM JURÍDICA

Tereza Afonso

tereza.afonso@gmail.com

Sílvia Araújo

saraujo@ilch.uminho.pt

Universidade do Minho, Portugal

Resumo

Direcionada para suprir a falta de tomada de consciência linguística (*language awareness*) pelos alunos de tradução, futuros tradutores, do funcionamento das línguas de trabalho e dos recursos linguísticos característicos das várias linguagens de especialidade, propomos uma abordagem didática, heurística, que parte do texto como unidade básica para um estudo contrastivo baseado num *corpus* paralelo. Através da abordagem proposta, os alunos analisam a macroestrutura (nível funcional-situacional) e microestrutura (nível formal-gramatical) do texto, adquirindo consciência das componentes lexicais, morfossintáticas, semânticas e pragmáticas dos textos escolhidos, das idiossincrasias de cada língua de trabalho usada num dado âmbito de especialidade, bem como dos fatores que influenciam as escolhas linguísticas de quem produz o texto, pois a linguagem não é neutra e tal é bem patente nas linguagens de especialidade. Como exemplo paradigmático de uma linguagem de especialidade altamente dependente das escolhas do poder público, dos valores que regem determinada sociedade em dado momento histórico, cultural e económico, foi escolhida a linguagem jurídica para ilustrar a abordagem que propomos. O *corpus* de estudo é composto de leis fiscais portuguesas. Trata-se do Código do Imposto sobre o Rendimento das Pessoas Singulares (IRS) e do Código sobre o Rendimento das Pessoas Coletivas (IRC), em vigor

desde 1988, na versão alterada pelos respetivos decretos-leis em 2010. A tradução para inglês dos referidos códigos figurava em 2016 no então Portal das Finanças da Direção-geral de Contribuições e Impostos, hoje Autoridade Tributária e Aduaneira. Para que a abordagem possa ser replicada noutros domínios de especialidade, é dada ênfase à metodologia de criação de um *corpus*, com indicação das ferramentas informáticas usadas para tal efeito.

Abstract

Aimed to overcome the lack of language awareness regarding the working languages and the linguistic features of the several specialized languages by translation students, future translators, this paper presents a didactic approach, considering the text as the basic unit, in order to teach them how to make a contrastive parallel corpus-based study. Through this approach, students will be able to analyze the macrostructure (functional-situational level) and microstructure (formal-grammatical level) of a text, therefore acquiring the awareness of lexical, morphosyntactic, semantic and pragmatic components of the chosen texts, as well as the idiosyncrasies of each working language used in a special domain. In addition, these students will also be able to uncover the factors that influence linguistic choices from the text producer, because language is not neutral, especially when it concerns specialized languages. In order to illustrate the application of this approach, legal language has been chosen as a paradigmatic example of a specialized language, which is highly dependent on the choices of the political power regarding the values that rule a given society, at a given historical, cultural and economic moment. The study corpus is composed by Portuguese tax laws, more precisely by the IRS and IRC Codes (tax law on the income of natural persons and collective persons), in force since 1988, in the version amended by the respective decree-laws in 2010. The English translation of these codes was online in 2016 in the latter *Portal das Finanças* of the *Direção Geral de Contribuições e Impostos* (General Directorate of Taxation and

Finances), renamed as *Autoridade Tributária e Aduaneira* (Tax and Customs Authority). To replicate this approach to other specialized domains, the methodology for creating a corpus and the computing tools used for this purpose are the focus of this paper.

Palavras-chave: linguagem de especialidade, linguagem jurídica, didática da tradução, consciência linguística, tecnologia, linguística de *corpus*

Keywords: Specialized language, legal language, didactics of translation, linguistic awareness, technology, corpus linguistics

1. Introdução

Para apreensão das características linguístico-estilísticas mais salientes da linguagem especializada, a compilação de textos paralelos constitui uma mais-valia em termos de autoajuda para tradutores (Tagnin, 2002; Granger *et al.*, 2003). Embora alguns autores, como Gellerstam (1996), alertem para os efeitos potencialmente nefastos de se expor os aprendentes de uma língua estrangeira às idiossincrasias típicas das traduções, pretendemos mostrar de que modo a linguística baseada em corpora (paralelos) contribui para o desenvolvimento (meta)linguístico dos alunos (Frankerberg-Garcia *et al.*, 2011). De facto, ao aprender a "ler" concordâncias bilingues, os alunos são necessariamente incitados a estabelecer comparações intra e interlinguísticas que reforçam o diálogo entre a língua materna e as línguas estrangeiras. Esta abordagem que Bernardini (2002) designa por *Corpus-aided Discovery Learning* permite, como veremos, a identificação e aquisição de particularidades sintáticas, pragmáticas, terminológicas e estruturas específicas de um domínio de especialidade. No âmbito do presente artigo, é nossa intenção apresentar uma metodologia

que ajude os alunos a compilar e analisar as características macro e microlinguísticas de corpora paralelos, constituídos a partir de textos de especialidade.

A metodologia que propomos tem por base a exploração de um *corpus* paralelo de textos legislativos portugueses, concretamente do Código do Imposto Sobre o Rendimento das Pessoas Singulares (IRS) e do Código do Imposto Sobre o Rendimento das Pessoas Coletivas (IRC) e a respetiva tradução para língua inglesa[1]. O trabalho foi desenvolvido em sala de aula no âmbito da unidade curricular *Seminário de Orientação e Profissionalização*, como preparação para o relatório de estágio/dissertação no ano de 2016[2]. Após comprovado sucesso, ma abordagem pedagógica semelhante está a ser aplicada nas unidades curriculares de *Linguística de Corpus* e *Seminário de Dissertação e Profissionalização*. Com este método pedagógico, pretendemos aproximar os alunos dos textos de especialidade, levando-os a refletir sobre os fenómenos linguísticos nas línguas de partida e de chegada, bem como a enquadrar tais fenómenos no correspondente campo temático e, sendo o caso, a evidenciar os aspetos sociais, históricos e culturais que possam ter relevância para efeitos de tradução. Começaremos por contextualizar a nossa proposta. De seguida, abordaremos a questão da necessidade da tomada de consciência dos fenómenos linguísticos pelos alunos de tradução, passando, depois, a expor a nossa proposta didática. Tomaremos algum tempo com a metodologia de compilação de um *corpus* para, seguidamente, apresentarmos a abordagem heurística que é aplicada nas aulas e nos debruçarmos sobre o caso de estudo de linguagem jurídica. Finalmente, apresentaremos as considerações finais.

[1] A tradução do Decreto-lei 442-A/88, de 30 de novembro (CIRS) e do Decreto-Lei n.º 442-B/88 (CIRC) datada de julho de 2010 com inclusão das alterações introduzidas nos respetivos Códigos pela Lei n.º 12-A/2010 (PEC 3) esteve a cargo de William Cunningham, consultor fiscal da Delloite Portugal. Os Códigos de IRS e IRC, assim como a Lei Geral Tributária (LGT), estiveram disponíveis em versão bilingue no sítio da Autoridade Tributária e Aduaneira, designado por Portal das Finanças, no separador *Portuguese Tax System*.

[2] O referido trabalho que incidiu sobre o fenómeno da modalidade foi alvo de uma comunicação com o título "Corpus-based analysis of modality in Portuguese-English legal texts" no primeiro congresso From Legal Translation to Jurilinguistics: Interdisciplinary Approaches to Study of Language and Law, que decorreu em Sevilha na Universidad Pablo de Olavide, nos dias 27 e 28 de outubro de 2016.

2. Contextualização

O objetivo mais amplo, que enquadra a proposta didática de que daremos conta seguidamente, prende-se com a tomada de consciência dos fenómenos linguísticos e com a aprendizagem decorrente da análise contrastiva baseada em *corpus*. Segundo Carter (2003, p. 64), a consciência linguística (*language awareness*)[3] refere-se ao desenvolvimento nos alunos de uma maior consciência e sensibilidade às formas e funções da linguagem. Para o mesmo autor, esta consciencialização carece de uma abordagem holística, baseada no texto como unidade básica, que comprove e demonstre como o uso da língua, não só não é neutro, como pode esconder ou salientar motivações de natureza social ou ideológica[4]. Ora, como afirma Faber (1998, p. 9): "*Such awareness is obviously a major asset for any foreign language (FL) learner, but for translation students, it is a vital necessity*», daí a nossa proposta didática assentar na construção de um *corpus* de estudo paralelo[5] e na sua exploração linguística.

Como explica Sardinha (2000, p. 357), "a Lingüística de Corpus é uma perspectiva, isto é, uma maneira de se chegar à linguagem", sendo um corpus (corpora, no plural):

> *um conjunto de dados linguísticos (pertencentes ao uso oral ou escrito da língua, ou a ambos), sistematizados segundo determinados critérios, suficientemente extensos em amplitude e profundidade, de maneira que sejam representativos da totalidade do uso linguístico ou de algum de seus âmbitos, dispostos de tal modo que possam ser processados por computador, com a finalidade de propiciar*

[3] O conceito de *language awareness* é geralmente associado ao movimento *British Language Awareness Movement*) e James, Garret (1991) e Hawkins (1992, 1999) sendo que a consciencialização do conhecimento linguístico teria de primeiro ser despertada nos professores (Cf. Casanova 2005, p.18)

[4] A este propósito, relembramos Rodrigues Lapa (1984, pp. 189-190): "Na linguagem oficial usa-se a voz passiva ou voz reflexa com valor de passiva, porque as determinações legais dirigem-se a uma massa passiva orientada superiormente por um órgão activo, que se adivinha sempre presente: o Estado. Assim se justifica o carácter impessoal dessa linguagem, para a qual valem, mais que as pessoas, os actos praticados por elas. […]. Do que fica exposto, conclui-se que o emprego da voz activa, passiva e reflexa não se faz às cegas. Há razões delicadas que impõem o seu uso, conforme as circunstâncias. Quem possui o sentimento da língua dificilmente se enganará nessa manipulação dos ingredientes do estilo».

[5] Conjunto de textos constituído de originais e correspondentes traduções, por oposição aos corpora comparáveis que comportam textos do mesmo género textual nas respetivas línguas.

> *resultados vários e úteis para a descrição e análise. "* (Sanchez 1995, pp. 8-9 *apud* Sardinha 2000, p.33)[6]

A abordagem mais comum é a análise baseada em *corpus* (*a corpus-based approach*), ou seja, o *corpus* fornece evidências e exemplos de padrões de uso da linguagem em análise[7], o que permite através de dados quantitativos complementar e dar substância à análise qualitativa. A exploração de um *corpus* por meio de ferramentas linguísticas tornou-se, entretanto, possível e acessível graças às novas tecnologias e ao desenvolvimento da Ciência da Computação. Aliás, a Linguística de *Corpus* recebeu um novo e grande fôlego desde os anos 90 até agora, possibilitando uma visão dos fenómenos linguísticos nunca concebida. Por outro lado, as análises linguísticas assumem especial relevância no que toca à linguagem de especialidade[8], pois "a especificidade das linguagens especializadas se expressa principalmente pela freqüência de uso de determinados recursos lingüísticos, comprováveis com o auxílio de métodos de estatística linguística» (Hoffmann, 2004, p. 81). Embora não haja uma definição única do conceito de linguagem de especialidade (Cabré, 1999, p. 61), é visível nos textos ligados a determinada área ou profissão a utilização de uma linguagem com rasgos próprios, desenvolvidos para acorrer a necessidades expressivas concretas dos falantes e à natureza das situações de comunicação por estes vividas. Pelo exposto, foi necessário encontrar uma metodologia adequada aos alunos de tradução que aliasse a teoria à prática, em linha com a era digital em que vivemos, e que preparasse os futuros tradutores para os desafios do século XXI.

[6] Ou, numa formulação mais sucinta: *"a body of language representative of a particular variety of language or genre which is collected and stored in electronic form for analysis using concordance software»* (ESRC/CASS, 2013, p. 5).
[7] Distinguindo-se da *corpus-driven approach*, em que os constructos linguísticos emergem do próprio *corpus*.
[8] Entendemos preferível a designação linguagem de especialidade para nos referirmos aos tecnoletos em lugar de língua de especialidade, usando a dicotomia língua geral – linguagem de especialidade, uma vez que as linguagens de especialidade comungam dos recursos linguísticos da língua comum numa relação de código e subcódigo (Cabré, 1999, p.58-59) e não de sistema e subsistema. Neste sentido, cf. Barros (2004, p. 43).

3. Promover a consciência metalinguística em domínios de especialidade

A tradução, apesar de ser um saber predominantemente operativo (Hurtado Albir, 2016, p. 25), carece de reflexão. Em especial, num domínio de especialidade, é fundamental a familiarização com o discurso utilizado pelos especialistas e profissionais da área, bem como um enquadramento que explica as opções tomadas por quem produz o texto original. Como afirma (Elena, 2008, p. 153): "El conocimiento textual intuitivo es la primera etapa de aproximación a un texto y, por tanto, la base sobre la que se construye el conocimiento textual científico propio de un experto.» Guiados pelo docente, os alunos têm a possibilidade de aprofundar conhecimentos e desenvolver as competências necessárias para analisar e traduzir os diferentes textos de especialidade, tendo em conta as características morfossintáticas, lexicais, terminológicas, fraseológicas, semânticas e pragmáticas da cultura de partida e da cultura de chegada.

4. Proposta didática de criação e exploração de um *corpus*

Apresentamos abaixo uma representação da proposta didática que temos vindo a testar, desde 2016, na formação dos alunos de tradução.

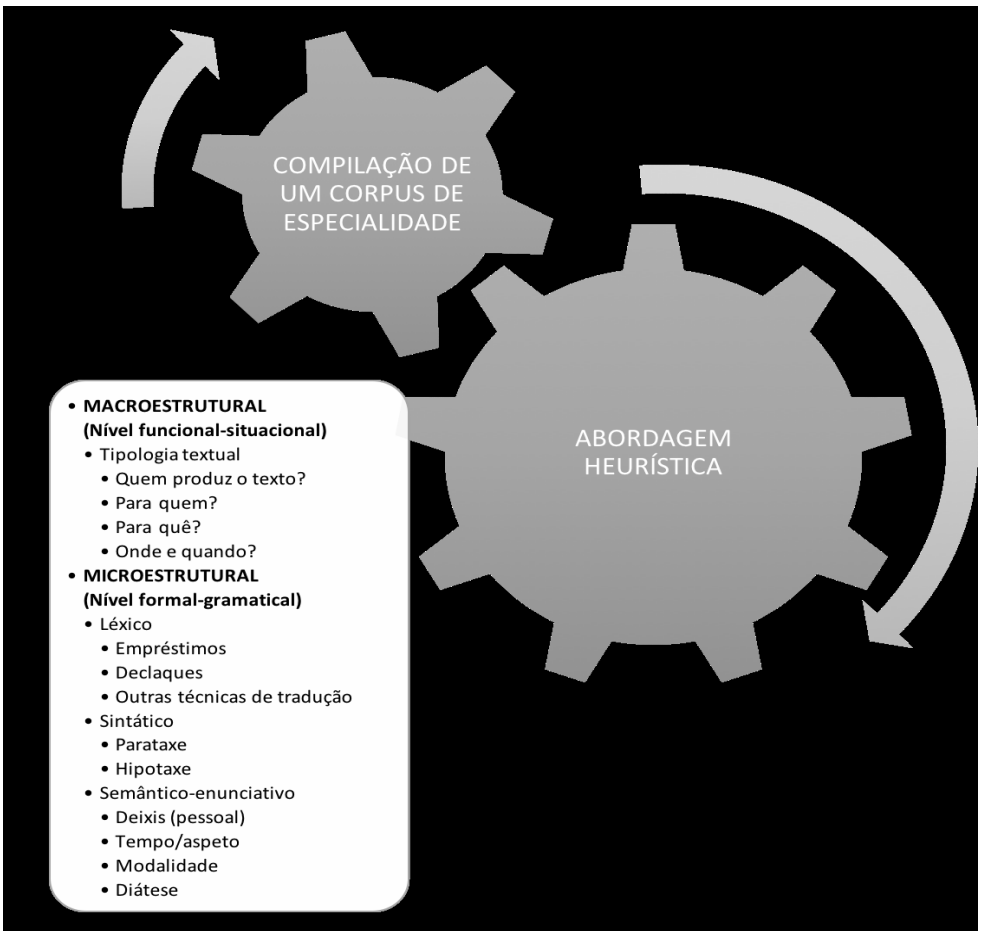

Figura 1. Proposta didática para criação e exploração linguística de um *corpus*

Como se pode ver acima, a proposta didática que preconizamos divide-se em duas etapas: a criação de um *corpus* de especialidade em formato TMX (*Translation Memory eXchange)* e a abordagem heurística para exploração linguística desse *corpus*. Esse *corpus* ou memória de tradução, que inclui os textos bilingues alinhados, tem um formato baseado em linguagem XML (*Extensible Markup Language*). É este formato que permite a partilha e a gestão das memórias de tradução. As memórias de tradução são frequentemente conhecidas por TMX (Almeida & Simões, 2007), confundindo-se o formato com o próprio ficheiro, à semelhança do que se passa com os ficheiros *Excel* ou *Word*.

A proposta que apresentamos tem os seguintes objetivos didáticos:

1) possibilitar um primeiro contacto com a Linguística de *Corpus*;

2) dotar o aluno de competências que lhe permitam, no futuro, criar os seus próprios recursos de tradução;
3) consciencializar o aluno da importância da seleção de fontes de informação fiáveis e credíveis que o ajudem a fazer face às encomendas de tradução e aos requisitos das mesmas, com qualidade;
4) exercitar a consciência (meta)linguística e fomentar um processo mental através de um guião que leve o aluno a explorar os diferentes níveis de um texto de especialidade.

4.1. Metodologia de criação de *corpus* paralelo

Primeiro, os alunos começam por familiarizar-se com a Linguística de *Corpus* e as suas múltiplas potencialidades por meio de textos paralelos. Cientes do efeito *translationese*[9], os textos paralelos não perdem a sua valia no que respeita ao ensino e aprendizagem de línguas e tradução, recomendando-se mesmo que os futuros tradutores comecem a alinhar textos para alimentar as ferramentas de tradução, conhecidas como CATtools. Um *corpus* paralelo prototípico, com alinhamento à frase, possibilita inúmeras aplicações, como assinalam BAKER *et al.* (2006: 127), entre elas a comparação lexical ou gramatical, as caraterísticas linguísticas dos textos traduzidos, o desenvolvimento de ferramentas de tradução automática e a integração em ferramentas de tradução assistida por computador (*CATtools*). Assim, o *corpus* serve de memória de tradução (MT)[10], ou seja, uma base de dados que armazena segmentos (frases, parágrafos ou unidades textuais como cabeçalhos, títulos ou elementos em uma lista) ou bitextos. Na senda de Zanettin (2012, p. 169), uma MT é um tipo de corpus

[9] Na definição de Santos (1998, p. 12, nota 18): "o desvio em relação à língua de destino que acontece em textos traduzidos, devido à interferência (inconsciente) da língua de origem (da sua gramática ou do seu léxico». Também Olohan (2004, p. 90): "*the term 'translationese' is a common description for translated language that appears to be influenced by the source language, usually in an inappropriate way or to an undue extent*».
[10] Ou em inglês, *Translation Memory* (*TM*).

paralelo. A noção de bitexto, termo cunhado por Harris em 1987 (*apud* Melby *et al.*, 2015, pp. 419-424), remete-nos para a mente do tradutor que, durante a prática da tradução, "lamina" o texto em segmentos para o poder traduzir. Cada segmento do Texto de Partida (TP) fica mentalmente ligado ao correspondente segmento do Texto de Chegada (TC) formando uma unidade do processo de tradução (unidade cognitiva). Este processo descoberto por via da Psicolinguística foi depois aplicado à tecnologia da tradução, designadamente às *CATtools*, à tradução automática e, ainda, às ferramentas de análise linguística. Mais tarde, os alunos, se assim o pretenderem, poderão trabalhar com corpora comparáveis. Os corpora paralelos podem ser integrados nas ferramentas de apoio à tradução através da sua exportação como MT. Assim, além de objeto de estudo, os corpora criados podem ser usados em contexto profissional.

Antes de proceder à análise linguística do *corpus* paralelo segundo a abordagem heurística que adiante apresentamos, os alunos têm de construir o *corpus* que pretendem estudar, selecionando os textos de especialidade. Atualmente, com a facilidade de busca na Internet[11], as escolhas recaem sobre textos que se encontram online em formato .pdf ou publicados em websites. No segundo caso, é necessário recorrer a ferramentas de *Web Scraping* (*Web Scraper*[12], *DownThemAll*[13], entre outros) para descarregar um website, que pode ser bi- ou multilingue. É raro, embora possível que os textos tenham de ser digitalizados e tratados por uma ferramenta que faça o reconhecimento ótico dos caracteres (OCR - *optical character recognition*). Sendo o caso, há ferramentas pagas ou gratuitas que possibilitam o reconhecimento dos carateres, embora quase sempre seja necessário efetuar correções manualmente.

[11] Zanettin (2002, p. 239): "*The WWW is the single largest existing repository of electronic texts, and has recently attracted the attention of researchers involved in translator training as a suitable source of texts for the creation of "disposable corpora". These are small, specialized corpora created* ad-hoc to serve the needs of the translator for a specific translation project, and their value lies not only in their analysis but even more so in their creation.

[12] *https://webscraper.io/*

13 https://www.downthemall.net/

Etapas da compilação do *corpus*

Conforme a figura que se segue, da construção de um *corpus* até à análise linguística com recurso à ferramenta *Sketch Engine,* existem seis fases.

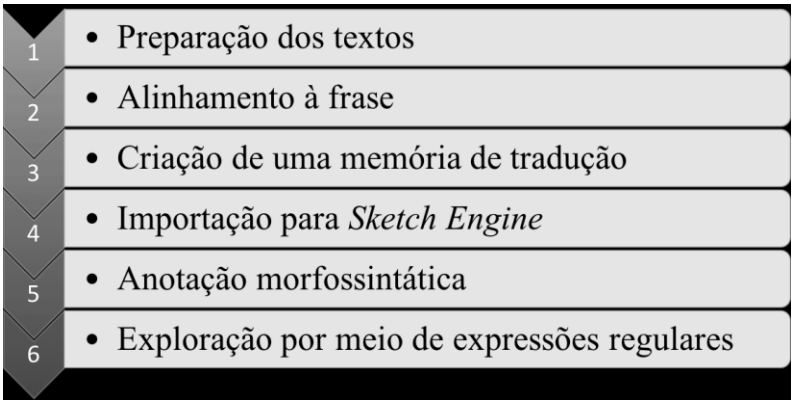

Figura 2 - Preparação e análise linguística de uma memória de tradução

Etapa 1: preparação dos textos

No caso escolhido, os textos encontravam-se disponíveis em formato pdf no portal das Finanças da então Direção geral de Contribuições e Impostos. Contudo, cada página apresentava duas colunas: o original em português e a tradução em inglês. Por conseguinte, foi necessário converter os documentos pdf em *Word*[14] e, de seguida, colocar as duas línguas em documentos separados de forma a passar à etapa seguinte. Cabe ao aluno encontrar as soluções que sejam adequadas à preparação dos textos do *corpus*, existindo uma panóplia de recursos gratuitos que o poderão ajudar nessa tarefa e podendo contar sempre com o apoio do docente.

[14] Foi usado o software gratuito Smallpdf

Etapa 2: alinhamento à frase

O alinhamento é uma noção intimamente ligada à segmentação em bitextos, isto é, à divisão do TP e do TC em unidades de tradução com o propósito de encontrar relações de equivalência ou correspondência, como explica Ahrenberg (2015, p. 395). O TP e o TC são postos em relação a fim de que possam ser estabelecidas concordâncias bilingues. O alinhamento mais comum é aquele em que a frase é tomada como unidade mínima. O alinhamento é feito automaticamente por programas criados para o efeito, o que não invalida a posterior revisão e edição manual. Atualmente, a maioria das ferramentas de apoio à tradução existentes no mercado possibilita o alinhamento de textos para criação de recursos de tradução. Por opção, no caso em apreço, recorremos ao *LF Aligner*, software gratuito, de código aberto[15] e multiplataforma (*Windows*, *OS X*, Li*n*ux) que permite construir memórias de tradução. Há sempre, em maior ou menor medida, alguma necessidade de edição manual de segmentos. Um dos exemplos típicos, que também sucede nas ferramentas de tradução, tem a ver com o facto de o programa assumir o ponto final como fim da frase, o que nem sempre é verdadeiro. Nos Códigos de IRS e IRC deparámo-nos com essa situação devido à abreviatura de artigo (art.º), divisão típica dos textos legislativos. Isso implicou uma correção manual em *Excel*.

Etapa 3: ficheiros em formato TMX

Ter o ficheiro em formato TMX permite que este funcione como memória de tradução, através da sua integração numa *CATtool*, assim como o seu processamento por ferramentas linguísticas denominadas concordanciadores. Este formato permite também a partilha de MT entre tradutores ou entre cliente e tradutor. Os concordanciadores consistem

[15] Disponível em https://sourceforge.net/projects/aligner/

em programas informáticos que encontram automaticamente concordâncias na sequência de instruções dadas pelo utilizador, construindo listas de ocorrências de uma palavra ou expressão procurada num *corpus* mediante o chamado *KWIC* (*Key Word In Context*). Portanto, uma concordância é um índice de todos os contextos em que uma palavra surge num *corpus* (Zanettin, 2015, p. 437). Um concordanciador fornece ao utilizador, além do número de ocorrências, informação sobre o contexto em que é utilizada a palavra ou expressão procurada. Neste aspeto, podemos estabelecer um paralelismo entre os dicionários e os corpora (monolingues e bilingues), por via das concordâncias que deles obtemos: ambos podem ser consultados para esclarecimento sobre o significado e uso de uma palavra mediante a interpretação do seu contexto; os corpora paralelos podem funcionar como dicionários bilingues (Zanettin, 2015, p. 440). No entanto, acima de tudo, um *corpus* paralelo pode fornecer informação sobre as estratégias de tradução utilizadas nos casos em que não existe equivalente direto (Zanettin, 2015, p. 441), como seja a possibilidade de resolver dúvidas linguísticas (por exemplo, regências verbais), de entender palavras que não estão dicionarizadas, de abarcar significados próprios de um dado domínio de especialidade, em que os recursos são escassos ou de reduzida qualidade. Por outro lado, ao poder criar corpora específicos para um projeto de tradução e extrair a terminologia e fraseologia que caracterizam as linguagens de especialidade[16], o tradutor ganha autonomia na produção de recursos essenciais à sua profissão.

[16] Um ficheiro em *Excel* com duas colunas (TP+TC), desde que o alinhamento correto tenha sido assegurado, pode ser convertido em formato TMX, embora tal conversão não seja, habitualmente, necessária, pois muitas *CATtools* e concordanciadores aceitam ficheiros em *Word*, *Excel* ou *Bloco de Notas* para criação de MTs ou corpora paralelos. Se a situação se colocar, há programas gratuitos ou pagos que solucionam a questão. Existem, de facto, aplicações de criação de memórias de tradução online, tais como *prompsit* (http://aplica.prompsit.com/pt/tmx), *Align Assist* (http://felix-cat.com/tools/align-assist/) ou *YouAlign* (https://youalign.com/).

Etapa 4: importação para o *Sketch Engine*

Atualmente, os concordanciadores apresentam-se como ferramentas robustas, que dão informações estatísticas sobre a frequência e posição das palavras no *corpus*, bem como fornecem listas de candidatos a termos. A escolha do *Sketch Engine*[17] deveu-se ao caráter intuitivo da ferramenta, um software de gestão de corpora e de análise textual. A ferramenta funciona online e é um software proprietário, embora exista um período experimental gratuito de um mês. Apesar de a empresa detentora do produto ter tornado o acesso às licenças de utilização mais acessível através de protocolos de cooperação com universidades, o preço pode constituir um entrave ao uso desta ferramenta. Duas grandes vantagens podem, no entanto, ser assinaladas em comparação com outros concordanciadores[18]: a integração de um anotador (ou etiquetador) morfossintático, cuja relevância será explicitada na fase seguinte, e de um *corpus* de referência, fundamental para elaborar listas de candidatos a termo. As figuras que se apresentam correspondem à interface "clássica" do *Sketch Engine*, ainda disponível para utilizadores conservadores. A importação de corpora criados pelo tradutor pode ser feita em TMX ou *Excel* e é um passo que não oferece dificuldades[19].

[17] Disponível em https://www.sketchengine.eu/

[18] Por exemplo, o *AntConc*, software gratuito disponível em http://www.laurenceanthony.net/software.html

[19] O programa aceita ainda outros formatos de ficheiros como .txt ou .doc, que podem ser carregados usando as opções *Parallel corpora* ou *Comparable corpora*, visíveis no lado esquerdo da figura abaixo.

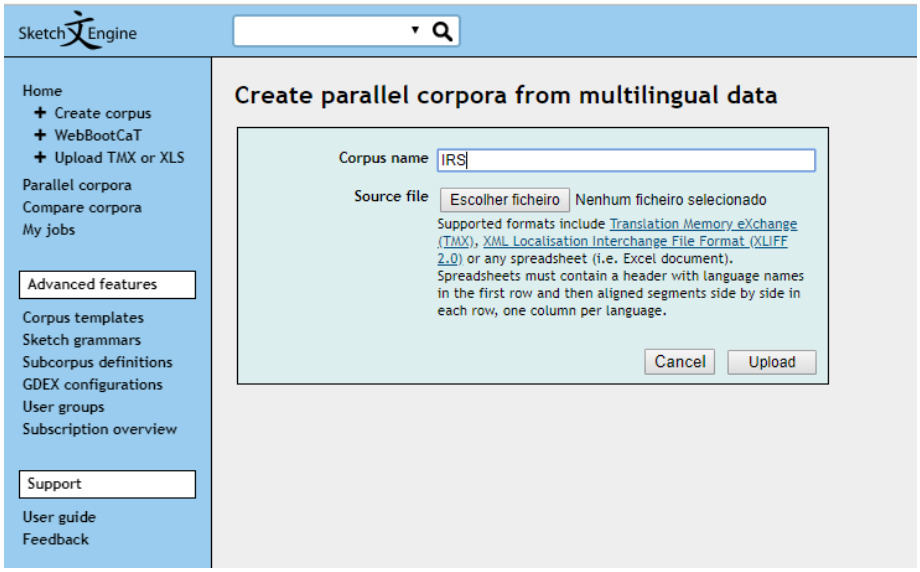

Figura 3 - Importação do ficheiro em *Excel* para a ferramenta *Sketch Engine*

Etapa 5 – Anotação morfossintática

A anotação morfossintática do *corpus* consiste em etiquetar cada palavra com informação gramatical. Também chamada anotação *Part-of-Speech* (POS), este processo automático, além de associar cada item lexical a uma categoria gramatical, inclui, ainda, a lematização, quer dizer, a determinação da forma canónica (lema) de cada palavra. Através destas etiquetas será possível processar o *corpus*, o mesmo é dizer, dar instruções à ferramenta de análise linguística para que esta, de forma rápida e automática, localize e devolva as ocorrências que pretendemos observar. A seleção do etiquetador *TreeTagger*[20], que está integrado no *Sketch Engine,* é o passo que se segue ao carregamento do *corpus*.

[20] Disponível em http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

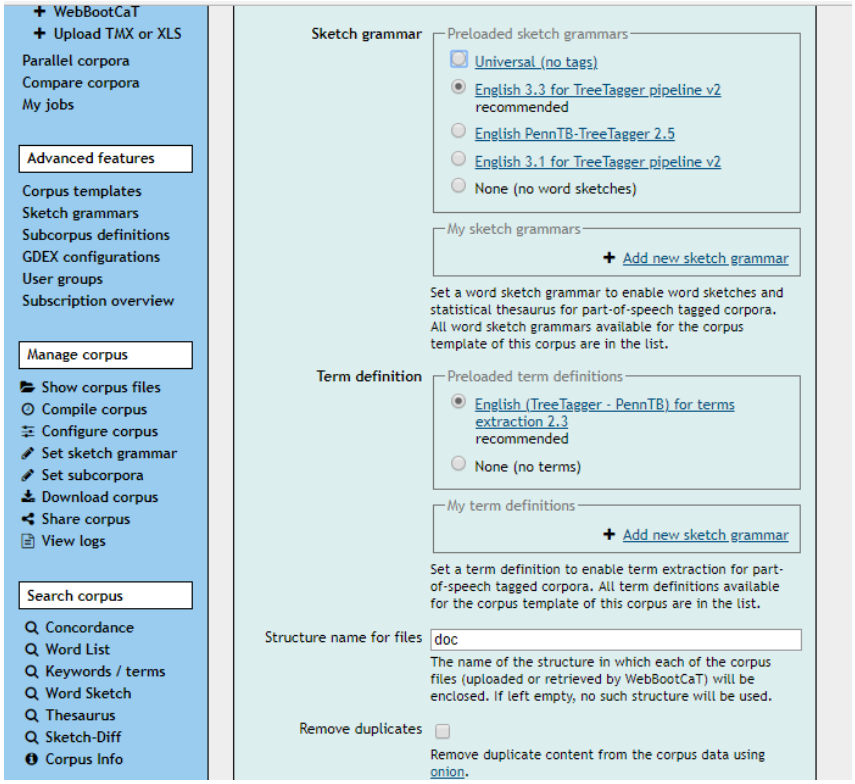

Figura 4 - Seleção do anotador

Etapa 6: Exploração do *corpus*

Depois de etiquetar a memória de tradução com anotação gramatical por forma a realizar pesquisas mais avançadas através da linguagem de consulta CQL (*Corpus Query Language*)[21] (Kilgariff *et al.*, 2014), os alunos são levados a explorar essa MT através da formulação de expressões regulares. Ao mobilizar as etiquetas *word, lemma* e *tag (pos)*, os alunos conseguem navegar pelos bitextos em busca de diferentes marcadores linguísticos, nomeadamente marcadores diatéticos, modais ou temporo-aspetuais. Ao introduzir, por exemplo, a fórmula [lemma="ser.*"] [tag="V.*"], é possível extrair todas as frases passivas (curtas ou longas) existentes no documento original (em português) e verificar se a diátese passiva dessas frases é mantida (ou não) aquando da passagem para a língua inglesa.

[21] https://www.sketchengine.eu/documentation/corpus-querying/

4.2 Abordagem heurística para exploração linguística do *corpus*

A análise de um texto depende dos objetivos específicos de quem o analisa. Os alunos são encorajados a analisar os textos (original e respetiva tradução) a dois níveis: a nível macro (aprofundamento do conhecimento das características de diferentes tipos de texto) e microestrutural (mecanismos linguísticos de coesão textual).

No primeiro plano (macroestrutural), os alunos procuram classificar os textos originais recorrendo ao separador intitulado "Tipologia Textual" disponibilizado no *Dicionário Terminológico* (DT) adotado pelo Ministério da Educação. Segundo este Dicionário, "os textos, para além das propriedades fundamentais da textualidade, apresentam estruturas verbais peculiares, semânticas e formais, e marcas pragmáticas que possibilitam a sua classificação em tipos ou géneros[22]. As características dos tipos ou géneros constituem indicadores importantes para a produção e para a interpretação dos textos[23]. Nesta etapa, trata-se, por conseguinte, de analisar a macroestrutura do texto (nível funcional-situacional), enquadrando os textos do *corpus* de estudo na tipologia textual, identificar quem produziu o texto, quem é o seu destinatário (ou destinatários), qual o intuito com que o texto foi produzido, onde e quando foi produzido. Na macroestrutura, integramos ainda a identificação das secções presentes no texto e a relação hierárquica entre elas. Esta é uma fase em que se cria uma espécie de "bilhete de identidade" dos textos escolhidos pelo aluno. Assim fica demonstrado, de forma imediata, o peso da seleção de determinada macroestrutura em detrimento de outra aquando da produção do texto. Na análise microestrutural (nível formal-gramatical), os alunos procedem a uma análise linguística contrastiva, debruçando-se sobre os aspetos lexical, sintático e semântico-enunciativo. Ao

[22] Esta é a classificação seguida, embora se reconheçam outras tipologias, como é o caso da proposta por Hurtado Albir (2016, pp. 242, 636), para quem os tipos textuais correspondem ao agrupamento de textos segundo a sua função – expositiva, argumentativa, instrutiva, e os géneros como agrupamentos de textos segundo a forma convencional e a situação de uso. Esta é a classificação seguida por Afonso (2016, pp. 35-38).
[23] *Dicionário Terminológico*, disponível em: http://dt.dgidc.min-edu.pt/ (consultado em 14/07/2019).

nível sintático, trata-se de indicar quais os tipos de construção sintática mais frequentes no original e quais as técnicas de tradução (Hurtado & Molina, 2002) utilizadas para traduzir essas construções. Ao nível semântico-enunciativo, os alunos analisam os textos no que diz respeito às principais categorias gramaticais que afetam o verbo, a saber, o tempo-aspeto (Campos *et al.*, 1991, 1997), a modalidade (Oliveira, 2003; Valentim, 2008) e a diátese verbal (Peres & Móia, 1995; Estrela, 2011).

4.2.1 Aplicação da abordagem heurística ao *corpus* paralelo português-inglês CIRS e CIRC

Sem grandes delongas sobre a linguagem ou as linguagens do Direito[24] ou sobre o que o que é ou não um texto jurídico, é inquestionável que a lei enquanto texto normativo assume, nos sistemas jurídicos de cariz romano-germânico, a expressão máxima da regulação das relações sociais dos cidadãos pelo Estado. Por essa razão, diz Borja Albi (2000, p. 11):

> *Se entiende por lenguaje jurídico el que se utiliza en las relaciones en que interviene el poder público, ya sea en las manifestaciones procedentes de este poder (legislativo, ejecutivo o judicial) hacia el ciudadano, o en las comunicaciones de los ciudadanos dirigidas a cualquier tipo de institución. Y también, naturalmente, el lenguaje delas relaciones entre particulares con transcendencia jurídica (contratos, testamentos, etc.).*[25]

[24] A linguagem do legislador, a linguagem dos juízes, a linguagem dos notários, etc.
[25] É uma definição que nos satisfaz quando pensamos no plano interno dos Estados, mas que deixa de fora a linguagem jurídica utilizada entre Estados em organizações supranacionais, como a União Europeia, ou internacionais, como a ONU ou a OCDE.

4.2.2 Caracterização do *corpus* de estudo

A tabela seguinte mostra a composição e o tamanho do *corpus* com indicação do número de *tokens* (todas as ocorrências de uma palavra incluindo repetições) e do número de *types* (as palavras que são diferentes, não repetidas).

	IRS (PT)	IRS (EN)	Total	IRC (PT)	IRC (EN)	Total
tokens	37 243	36 992	**74 235**	48 928	47 723	**96 651**
types	31 174	31 025	**62 199**	41 792	40 732	**82 524**

Tabela 1 - Composição do *corpus* de estudo

Pelos dados apresentados, é possível verificar que o *corpus* cumpre seu propósito: a linguagem jurídica pertence a um único género (textos normativos), o que garante homogeneidade, e assegura a representatividade. Portanto, é uma amostra credível. Outro aspeto importante, que pesou na seleção dos textos, foi a fonte da informação e a identificação do tradutor, o que nos deu garantias sobre a compreensão do TP e a qualidade linguística (e, por sinal, terminológica)[26] do TC.

4.2.3. Análise macroestrutural

O *corpus* de estudo compõe-se de leis. Em consonância com a definição de linguagem jurídica proposta por Borja Albi (2000), também optámos pela tipologia da autora quanto

[26] William Cunningham, irlandês, viveu doze anos em Portugal. Em 2013, foi distinguido pelo Instituto de Direito Económico Financeiro e Fiscal (IDEFF) da Faculdade de Direito de Lisboa e a Revista de Finanças Públicas e Direito Fiscal como um dos "Senadores da Fiscalidade» - cf. https://goo.gl/9y78rR

aos textos jurídicos, porque entendemos que é mais adequada e pertinente para a organização mental de um tradutor jurídico[27].

Importa realçar que a tradução jurídica se distingue relativamente a outros tipos de tradução, porque "each legal system has its own language(s) and its own system of reference» (Šarčević, 1997, p. 230). Por essa razão, "law and legal language are system-bound, that is, they reflect the history, evolution and culture of a specific legal system» Cao (2007, pp. 23-24). No *corpus* de estudo, apesar do destinatário ser indefinido, internacional, a tradução de português para um inglês fortemente influenciado pelo sistema jurídico irlandês torna o texto próximo dos leitores familiarizados com o *common law* vigente nos países pertencentes à *Commonwealth*[28]. Portanto, a tradução não deixa de pôr relação duas línguas diferentes e dois sistemas jurídicos[29] diferentes (tradução interlinguística e intersistémica[30]). Por seu turno, dadas as especificidades da tradução jurídica Prieto Ramos (2009, 2011, pp. 14-16) propõe: 1) identificar os sistemas jurídicos envolvidos (coordenadas geográficas, ou seja, jurisdicionais e linguísticas); 2) identificar o ramo do Direito a que pertence o TP

[27] Esta tipologia agrupa os textos jurídicos em géneros, atende à situação discursiva, aos participantes no ato de comunicação, ao registo utilizado (geralmente, formal ou muito formal), ao modo (em regra, escrito ou escrito para ser lido), à finalidade e ao foco contextual; Borja Albi (2000, pp. 84-85) reparte os textos jurídicos por seis categorias:
1. Textos normativos (ex: Constituição, lei/*acts, statutes*);
2. Textos judiciais (ex: peças processuais, despachos, sentenças/*claims, judgments*)
3. Textos jurisprudenciais (ex: coletâneas de jurisprudência, *Law Reports*)
4. Textos doutrinais (ex: artigos científicos, manuais)
5. Obras de referência (ex: dicionários jurídicos)
6. Textos de aplicação do Direito;
a) Documentos privados (ex: contratos, testamentos/*contracts, legal letters*)
b) Documentos públicos (ex: certidões, escrituras públicas/*certificates, deeds*).

[28] Nota do tradutor.

[29] Aqui, sistema jurídico é usado como sinónimo de ordenamento jurídico.

[30] No caso da legislação europeia, os atos legislativos são publicados nas 24 línguas oficiais da União Europeia em versões com igual valor jurídico; tal só é possível após um processo que envolve tradução interlinguística, mas intrasistémica, uma vez o processo legislativo europeu corresponde a um único sistema. Poder-se-á compilar um *corpus* paralelo em que ambas as versões dos textos são originais, ainda que o processo legislativo tenha sido espoletado numa língua da UE, provável, mas não necessariamente inglês, passando por um processo complexo que envolve juristas-linguistas e que culmina na publicação no *Jornal Oficial da União Europeia* (JO). Entre outros exemplos paradigmáticos de tradução interlinguística e intrasistémica que costumam ser apontados estão as resoluções da ONU ou a legislação de Estados bilingues ou plurilingues como a Bélgica e a Suíça.

(coordenadas normativas e temáticas); 3) enquadrar o TP numa tipologia textual (coordenadas procedimentais, contextuais e discursivas).

Respondendo às perguntas do nível funcional-situacional, os textos de partida têm como autor o legislador português, que se dirige aos cidadãos portugueses para regular as relações fiscais dos contribuintes, pessoas singulares e coletivas. A versão de ambas as leis, em vigor desde 1988, remonta a 2010 com as alterações introduzidas pelo PEC (Programa de Estabilidade e Crescimento) aprovado pela Lei 12-A/2010 de 30 de junho, uma altura crítica para Portugal.

Acresce dizer que os códigos são leis de grande dimensão com uma estrutura característica.[31] Assim, depois do preâmbulo, o texto do código está organizado por Capítulos, Secções, Subsecções, Artigos e, sendo o caso, Parágrafos, Alíneas e Subalíneas.

4.2.4. Análise microestrutural

A possibilidade de análise em conformidade com a abordagem pedagógica proposta é demasiado exaustiva para ser apresentada no presente artigo. Referimos, apenas, a opção do legislador português pelo Presente do Indicativo, que evoca o tempo atual de quem lê independentemente da data de entrada em vigor da lei e das alterações que a mesmo possa ter sofrido. Em inglês, a opção de redação remete para o futuro com utilização de *Shall*, pelo que ocorre uma modulação na tradução. Outros aspetos poderão ser verificados na redação típica dos atos normativos, em conformidade com as regras de redação por que se rege o legislador português[32].

[31] Cf. *Guia prático de regras a observar na redação de actos normativos da Assembleia da República, (disponível em* https://www.parlamento.pt/DossiersTematicos/Documents/Reforma_Parlamento/guialegisticaformal.pdf).
[32] Cf. Colaço, L & Araújo, M.L (2008) e Resolução do Conselho de Ministros n.º 90-B/2015,https://dre.pt/home//dre/70961384/details/maximized?serie=I&dreId=70961381 [consultado a 14 de julho de 2019]

Em termos pragmáticos, a par da intenção de comunicação, há que considerar a força ilocutória dos textos legislativos dada a normatividade que lhes é inerente, torna apetecível a análise dos recursos usados para concretizar permissões, proibições ou obrigações, investigando assim, o papel dos auxiliares modais, como expressão e medida dessa força ilocutória.

Com um *corpus* relativamente pequeno, um *DiY-Corpus* nas palavras de Zanettin (2002), procurámos dar uma ideia das possibilidades de exploração e extração de informação mediante a aplicação desta abordagem heurística. Quanto maior conhecimento tiver o tradutor sobre o campo de especialidade que elege, maior confiança terá para enfrentar e resolver os desafios de tradução com que se deparar. Na tradução jurídica, é dever do tradutor traduzir Direito por Direito, ou seja, transpor a linguagem de especialidade do TP para a linguagem de especialidade do TC. Por isso, se exige um conhecimento profundo do funcionamento das línguas de trabalho e dos sistemas jurídicos que as enquadram.

5. Considerações finais

O percurso pedagógico acima descrito dispõe-se a ajudar os alunos a tomar consciência do processo de tradução, isto é, a desenvolver uma consciência linguística que lhes permita justificar as opções de tradução. Nesse sentido, a ferramenta *Sketch Engine* revela-se extremamente útil, porque permite que os alunos construam corpora paralelos de especialidade e procedam à sua análise, quer ao nível terminológico, quer aos níveis sintático e semântico-enunciativo. A Linguística de *Corpus* possibilita inúmeras aplicações em termos de Linguística (teórica ou aplicada), Estudos de Tradução e ensino de línguas. Ao fomentar a preocupação pelo rigor na observação, no raciocínio e no discurso metalinguístico (Duarte, 2008), os corpora participam, sem dúvida, de uma abordagem heurística que permite ver os fenómenos linguísticos a uma nova luz (Charaudeau, 2011). A compilação de corpora torna-

se igualmente produtiva para os alunos, que além dos conhecimentos adquiridos sobre um domínio de especialidade, ficam com uma MT pronta que podem usar profissionalmente, como também aprendem a ter autonomia relativamente à produção de recursos. Urge referir que todos os anos se afina a aplicação da abordagem didática, até porque o mundo está em rápida evolução e os alunos também são outros, com outros interesses e necessidades.

Referências bibliográficas

Afonso, T. (2016). *Tradução jurídica à luz da linguística de corpus*. Instituto de Letras e Ciências Humanas: Universidade do Minho.

Almeida, J.J., Simões, A. (2007) XML: TMX : processamento de memórias de tradução de grandes dimensões. In José Carlos Ramalho, João Correia Lopes & Luís Carriço (eds.), *XML: Aplicações e Tecnologias Associadas (XATA2007)* (FCUL, Lisboa, 15-16 de Fevereiro), Universidade do Minho, pp. 83-93. Disponível em: http://xata.fe.up.pt/2007/papers/7.pdf

Ahrenberg, L. (2015). Alignment. In Chan, S. (Ed.) *Routledge Encyclopedia of Translation Technology* (pp. 395-408). New York: Routledge.

Ançã, M. H. (2015). Revisitando a consciência linguística: apropriação do conceito por parte de futuros professores de Português. *Calidoscópio.* Vol.13(1), 83-91. Disponível em: https://goo.gl/z5DCTH

Baker, P., Hardie, A., & McEnery, T. (2006). *A glossary of corpus linguistics*. Edinburgh: Edinburgh University Press.

Barros, L.A. (2004). *Curso básico de terminologia*. São Paulo: Edusp.

Bernardini, S. (2002). Exploring new directions for discovery learning. In B. Kettemann & G. Marko (Eds.) *Teaching and learning by doing corpus analysis*. New York: The Edwin Mellen Press, 165-182.

Cabré, M.T. (1999). *Terminology: Theory, methods and applications* (Volume 1). Amsterdam: John Benjamins Publishing.

Campos, M.H.C. & Xavier, M.F.. (1991). *Sintaxe e Semântica do Português*. Lisboa: Universidade Aberta.

Campos, M. H. C. (1997). *Tempo, aspecto e modalidade: estudos de linguística português*. Porto: Porto Editora.

Cao, D. (2007). *Translating Law. Topics in Translation*, Volume 33. Clevedon/Buffalo/Toronto: Multilingual Matters.

Carter, R. (2003). Language Awareness. *ELT journal*, *57*(1), 64-65. Disponível em: https://goo.gl/zdTLng

Casanova, I. (2006). *Linguística Contrastiva: O ensino da língua inglesa*. Lisboa: Universidade Católica Editora.

Charaudeau, P. (2011). Diz-me qual é teu *corpus*, eu te direi qual é a tua problemática. *Revista Diadorim*, 10, 1-23.

Colaço, L & Araújo, M.L. (2008). *Regras de LEGÍSTICA a Observar na Elaboração de Actos Normativos da Assembleia da República.* Lisboa: Divisão de Edições da Assembleia da República. Disponível em: http://www.asg-plp.org/upload/cadernos_tematicos/doc_160.pdf

De Groot, G.R. (1987). *La traduction juridique: The point of view of a comparative lawyer. Les Cahiers de droit*, 28(4), 793-812. doi: https://doi.org/10.7202/042842ar

Duarte, I. (2008). *O conhecimento da língua: desenvolver a consciência linguística.* Lisboa, Direção Geral de Inovação e Desenvolvimento Curricular.

Elena, P. (2008). La organización textual aplicada a la didáctica de la traducción. Quaderns: *revista de traducció*, 15, 153-167.

EMT COMPETENCE FRAMEWORK - 2017, disponível em https://ec.europa.eu/info/sites/info/files/emt_competence_fwk_2017_en_web.pdf

ESRC/CASS. (2013). *Corpus: Some key terms. CASS: Briefings. ESRC (Economic and Social Research Council) Centre for Corpus Approaches to Social Science (CASS)*, Lancaster University, UK. Disponível em: http://cass.lancs.ac.uk/?page_id=956

Estrela, A. P. (2011). A construção passiva: usos e desvios. In M. Teixeira, I. Silva, & L. Santos (Eds.), *Novos Desafios no Ensino do Português* (pp. 92-98). Santarém: Escola Superior de Educação de Santarém.

Faber, P. (1998). Translation competence and language awareness. *Language Awareness*, 7(1), 9-21. doi:https://doi.org/10.1080/09658419808667097

Frankerberg-Garcia, A., Flowerdew, L., Aston, G. (2011). *New Trends in Corpora and Language Learning*. London: Continuum.

Gellerstam, M. (1996). Translations as a source for cross-linguistic studies. In Karin Aijmer, Bengt Altenberg & Mats Johansson (eds.) *Languages in contrast: papers from a symposium on text-based crosslinguistic studies*. Lund Studies in English 88. Lund University Press, 53-62.

Granger, S., Lerot, J. & Petch-Tyson, S. (dir.) (2003). *Corpus based Approaches to Contrastive Linguistics and Translation Studies*. Amsterdam: Rodopi.

Harris, B. (1988). Bi-text, A New Concept in Translation Theory. *Language Monthly,* 54, 8–10.

Hawkins, E. (1992). Awareness of language/knowledge about language in the curriculum in England and Wales: An historical note on twenty years of curricular debate. *Language Awareness*, 1(1), 5-17.

Hawkins, E. W. (1999). Foreign language study and language awareness. *Language awareness*, 8(3-4), 124-142.

Heylen, K., & Steurs, F. (2014). Translating legal and administrative language: How to deal with legal terms and their flexible meaning potential. *Turjuman*, 23(2), 96-146.

Hoffmann, L. (2004). Conceitos básicos da lingüística das linguagens especializadas. *Cadernos de Tradução*, 17, 79-90.

Hurtado Albir, A. (2016). *Traducción y traductología*: *Introducción a la traductología*. 8ª Edição. Madrid: Cátedra.

James, C. & Garrett, P. (1991). *Language Awareness in the Classroom.* Harlow: Longman.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1(1), 7-36.

Melby, A., Lommel, A., & Morado Vazquez, L. (2015). Bitext. In Chan, S. (Ed.) *Routledge Encyclopedia of Translation Technology* (pp. 409-424). New York: Routledge.

Oliveira, F. (2003). *Modalidade e modo*. In MATEUS *et al.*, Gramática da Língua Portuguesa. 5ª edição revista e aumentada (pp. 243-272). Lisboa: Caminho.

Olohan, M. (2004). *Introducing corpora in translation studies*. London: Routledge.

Peres, J. & Móia, T. (1995). *Áreas Críticas da Língua Portuguesa*. Lisboa: Editorial Caminho.

Prieto Ramos, F. (2009). Interdisciplinariedad y ubicación macrotextual en traducción jurídica. *Translation Journal*, 13(4). Disponível em: https://archive-ouverte.unige.ch/unige:5078

Prieto Ramos, F. (2011). Developing legal translation competence: An integrative process-oriented approach. *Comparative Legilinguistics-International Journal for Legal Communication*, 5, 7-21. Disponível em: https://archive-ouverte.unige.ch/unige:16166

Rodrigues Lapa, M. (1984). *Estilística da Língua Portuguesa*. 11ª Edição. Coimbra: Coimbra editora.

Rodrigues, M. D. C. C. (2005). *Contributos para a análise da linguagem jurídica e da interacção verbal na sala de audiências*. Faculdade de Letras. Universidade de Coimbra. Disponível em: https://estudogeral.sib.uc.pt/handle/10316/714

SantoS, D. (1998). A relevância da vagueza para a tradução, ilustrada com exemplos de inglês para português. *Tradterm*, 5(1), 41-70. doi:https://doi.org/10.11606/issn.2317-9511.tradterm.1998.49774

Šarčević, S. (1997). *New approach to legal translation*. The Hague/London/Boston: Kluwer Law International.

Sardinha, T. B. (2000). Lingüística de corpus: histórico e problemática. *Delta*, 16(2), 323-36

Sardinha, T. B. (2004). *Lingüística de corpus*. São Paulo: Editora Manole.

Tagnin, S. E. O. (2002). Os Corpora: instrumentos de auto-ajuda para o tradutor. *Cadernos de tradução*, 1(9), 191-219.

Valentim, H. (2008). Modos gramaticais e modalidades - algumas particularidades do Português Europeu. In Anna Kalewska (org.) Diálogos com a Lusofonia. Actas do colóquio comemorativo dos 30 anos da Secção Portuguesa do Instituto de Estudos Ibéricos e Ibero-americanos da Universidade de Varsóvia (pp. 421-438). Instytut Studiów Iberyjskich i Iberoamerykańskich UW Warszawa.

Zanettin, F. (2002). DIY Corpora: The WWW and the Translator. In Maia, B.; Haller, J. & Urlrych, M. (eds.) *Training the Language Services Provider for the New Millennium*, Porto: Faculdade de Letras, Universidade do Porto, 239-248.

Zanettin, F. (2012). *Translation-Driven Corpora. Corpus Resources for Descriptive and Applied Translation Studies*. Manchester: St. Jerome.

Zanettin, F. (2015). Concordancing. In Chan, S. (Ed.) *Routledge Encyclopedia of Translation Technology* (pp. 437-449). New York: Routledge.