a folha

Boletim da língua portuguesa nas instituições europeias

http://ec.europa.eu/translation/portuguese/magazine

N.º 37 — Outono de 2011

Da revisão com leitura cruzada

Luís Seabra
Direcção-Geral da Tradução — Comissão Europeia

Primeiro estranha-se, depois entranha-se. Ouvi-o muitas vezes de colegas das instituições europeias, mas não só, quando em 1999 me juntei às fileiras da Direcção-Geral da Tradução (DGT) no edifício Jean Monnet. O Luxemburgo, garantiram-me, começa por ser um lugar estranho, mas passado algum tempo «entranha-se» e não é fácil largar. O homem é um animal de hábitos, sabemo-lo bem, e a resistência à mudança é uma das características primárias do ser humano. Respeitando, portanto, esta espécie de lei natural, resisti, como aliás grande parte dos colegas, à tradução com o Trados (TWB), ao desaparecimento das bibliotecas «nacionais» em favor de uma única biblioteca central e, para não me alongar muito e ir directamente ao que hoje aqui me traz, resisti à revisão com leitura cruzada.

Ora, de todas as práticas de revisão seguidas na DGT, a leitura cruzada é, para mim, a que oferece maiores vantagens e produz melhores resultados, que se reflectem não só na qualidade do texto final, mas também na qualidade da relação profissional entre colegas. Para dissipar eventuais dúvidas, esta prática consiste no seguinte: o tradutor lê em voz alta a sua versão ao revisor, que segue o texto original em papel.

A vantagem maior e mais evidente decorre da presença dos dois colegas no mesmo gabinete, o que permite debater as dúvidas e encontrar de imediato soluções para elas e para eventuais problemas do texto original. Parece-me evidente que há uma diferença entre ler um texto a alguém em voz alta ou relê-lo (muitas vezes apressadamente) para o entregar com vista a uma revisão diferida, na qual o tradutor estará ausente. Ou seja, ao ler a própria versão em voz alta, para um ouvinte ali presente, o

tradutor irá forçosamente fazer algumas alterações ou adaptações que contribuirão para melhorar a inteligibilidade e fluidez do texto final.

Nem sempre as escolhas ou motivações do tradutor são claras para o revisor e todos ganhamos (não só em tempo!) se houver oportunidade de as expor e debater de viva voz, confrontando o original e a tradução. É verdade que, na forma mais habitual de revisão (em diferido), estas explicações também podem ser dadas, no momento em que se entrega o documento ao revisor. Porém, se este estiver ocupado com outras revisões, irão passar alguns dias ou mesmo semanas, o suficiente para que essas explicações se esfumem, sendo necessário repetir o que já foi dito e esquecido, o que representa uma perda de tempo.

Outra das vantagens é uma maior responsabilização do tradutor, com a correspondente satisfação profissional que dela pode retirar. Deste modo, deixa de ser possível entregar o texto ao revisor com uns «presentes» por resolver. E o revisor poderá então concentrar-se nos elementos essenciais da sua tarefa: detectar omissões e erros de interpretação ou contribuir para reformular frases menos conseguidas em português. Quem tem experiência de revisão — praticamente todos nós, portanto — sabe que a tentação é grande, que quando nos sentamos no lugar do revisor empunhando a caneta encarnada (ou verde, ou azul, ou mesmo um lápis, para o caso tanto faz), nada é mais fácil do que ceder e começar a alterar por ali fora, quanto mais não seja para não ficarmos com a sensação de não ter feito um bom trabalho... de revisão. Mas, na verdade, o autor e responsável final pela tradução que sai da DGT é o tradutor, pelo que é inútil e muito mais morosa a revisão que procura moldar a redacção do texto mais «ao gosto» do revisor. Não só mas também pelo tempo necessário para introduzir depois todas as alterações.

Outra das vantagens da leitura cruzada consiste, a meu ver, na possibilidade de reforçar a relação profissional entre certos colegas que de outra forma quase não têm oportunidade de trabalhar juntos.

«Mas eu não posso, és maluco, isto assim demora muito mais tempo!» É esta a desculpa habitual atrás da qual se esconde, com o rabo de fora, a resistência inicial de muitos colegas, escudando-se no (presumido) desperdício de tempo para nem sequer a experimentarem. Na verdade, conheço poucos colegas que não se tenham convertido depois de alguma experiência. Aproveitando os conhecimentos especializados de cada um e usufruindo dos muitos anos de experiência dos colegas mais velhos, é sempre mais benéfico debater as dúvidas e explicar as soluções encontradas directamente ao revisor do que encontrar-se frente às suas correcções, algumas vezes pouco compreensíveis ou mesmo indecifráveis (depende do grau de gatafunhanço), ainda que exista sempre possibilidade de «se não perceberes alguma coisa ou não concordares, não hesites em falar comigo» para desfazer dúvidas e perplexidades. Isto sim, representa uma perda de tempo: é preciso ir ao gabinete do revisor ou recorrer ao correio electrónico, esperar pela resposta, que às vezes mete réplica e tréplica, e etc., etc.

Ora, na maioria das vezes, tempo é exactamente aquilo que mais nos falta até ao prazo de saída do documento. Ou seja: não há tempo, não se faz. As dúvidas não se esclarecem, o tradutor decide introduzir as alterações ou não, podem nascer mal-entendidos (que se podem até eternizar) e uma coisa é certa: ficamos todos a perder por não partilharmos a nossa experiência e as opções que seguirmos naquele caso concreto e a qualidade final da tradução também sai prejudicada.

O nosso trabalho é tanto mais rico quanto maior for o número de colegas com quem trocamos ideias, experiências e dificuldades. Agora que o Departamento de Língua Portuguesa da DGT entrou numa fase de flexibilização total, na qual em boa hora todos podem ajudar todos, poderia até pensar-se na possibilidade de alargar a flexibilização também às revisões com leitura cruzada, desde que a proximidade geográfica assim o permitisse e que isto não traga mais burocracias inúteis e morosas, o que exclui à partida a cooperação entre o Luxemburgo e Bruxelas neste domínio (convenhamos que não é muito prático fazer leitura cruzada ao telefone ou em videoconferência), mas inclui as unidades 1 e 2 de Bruxelas, instaladas no mesmo edifício (Genève 1) em Evere. Fica feita a sugestão.

A (presumível) demora da leitura cruzada é ilusória. Senão vejamos: se, durante a revisão, o tradutor estiver a ler o texto directamente no computador e for introduzindo directamente as alterações, no final o documento está materialmente pronto, faltando só verificar os aspectos formais e proceder às «limpezas» necessárias (códigos do TWB, etc.). Por outro lado, parece-me que o revisor perde muito mais tempo se tiver de ler e comparar, sozinho, a versão original e a versão traduzida. E que desta forma poderá deixar escapar mais pormenores do que na leitura cruzada.

A maior dificuldade prática da leitura cruzada é conjugar a disponibilidade dos dois implicados de modo a poderem trabalhar juntos, sobretudo no caso de documentos mais longos ou de colegas teletrabalhadores. No entanto, se também aqui se usar de alguma flexibilidade, juntamente com uma boa dose de boa vontade, as soluções surgirão por si, afastando os obstáculos intransponíveis.

Luis.Seabra@ec.europa.eu

Da estratégia do curiosamente mal posicionado com *particules élémentaires pas vraiment nécessaires*

Luís Filipe PL Sabino
Antigo funcionário — Comissão Europeia; Comité Económico e Social Europeu-Comité das Regiões

[Texto adaptado a partir de artigo publicado no *Diário do Sul* (Évora), de 9.9.2011.]

Os usos linguísticos são curiosos. Desde há uns tempos o advérbio «sinceramente» acompanha todas as frases, declarações mais ou menos solenes, produzidas em qualquer ambiente. Assim, diz-se: sinceramente gostei deste filme; sinceramente não sei se vou; sinceramente o preço era esse, etc. Sinceramente, porquê o uso e abuso deste advérbio? Talvez seja porque, num contexto onde se mente bastante, onde amiúde não se diz o que se pensa — aliás, e já agora, poderá uma sociedade viver sem a mentira? Há sinceramente sérias dúvidas... de que pudesse subsistir uma sociedade assim (aliás, não consta dos livros de História!), sendo ainda certo que num ambiente social, económico e político onde as estatísticas adquiriram foros de cidadania, elas (estatísticas) figuram entre as categorias de mentiras mais difundidas (v. a este propósito, com interesse, o artigo publicado no *Financial Times* de 24.8.11, p. 9, sob o título «*On sex, lies and the pitfalls of overblown statistics*») —, há que acentuar que se está a dizer a verdade nua e crua, sincera, do fundo do coração, *ab imo pectore*, sem mentiras.

[No plano das mentiras históricas, é de recordar o massacre de Katyn (Rússia) de abril de 1940, quando os soviéticos assassinaram cerca de 22 000 prisioneiros polacos (das elites militares e civis), chacina que a propaganda comunista atribuiu à Alemanha nazi, de resto também esta especialista na tortura e na selvajaria: mentira mantida com diversas conivências até ao colapso da URSS. A propósito desta carnificina e de outros crimes nefandos cometidos pelos Nazistas e pela União Soviética na zona entre a Polónia central e a Rússia ocidental de 1939 a 1945, v. Timothy Snyder, *Bloodlands — Europe between Hitler and Stalin,* 523 p.p., editado por The Bodley Head, Londres 2010; e já agora, releia-se, nesta linha, André Gide — *Retour de l'URSS*.]

Mas, sinceramente, deixando de parte o sinceramente, há também outro termo que, por vezes a despropósito, se usa com frequência: é a palavra «curiosamente». Porquê dela abusar assim, coitadinha, quando às vezes nada há de curioso? Não há radialista, jornalista da TV ou da imprensa que a esse advérbio não recorra com dispensável assiduidade; eles ou elas (alguns/algumas: é preciso

não generalizar) até não o fazem por mal... porque até curiosamente dessa inutilidade não se apercebem: perdoai-lhes Senhor, Vós que tão tolerante sois!

Há ainda outro advérbio empregue a esmo: o «basicamente» que se põe no início de frases e que parece transmitir algo de definitivo e de profundo, o que assim não ocorre. É apenas mais um dispensável. Até me parece que esse basicamente vem suceder a um outro que ainda tem grande aceitação: o «fundamentalmente», que basicamente tem o mesmo significado e que, por via de regra, faz tanta falta como um leão à solta numa creche.

E o que me dizem do «estrategicamente»?! Qualquer caramelo lança mão do estrategicamente sem estratégia nenhuma, assim dizendo, v.g., que o presidente estava estrategicamente posicionado... (este verbo «posicionar» veio destronar o «colocar», passando-se a utilizar este em frases como «colocar questões», que muito gabiru, tendo-se por bem falante, prefere a «fazer perguntas», considerada esta, presumo, algo fatela. Nisto, também TV, radialistas & Cia. são useiros e vezeiros).

Outro termo que se tornou muito popular (o Vasco Pulido Valente — irado com o mundo em geral e não logrando salvar a sociedade — diria «muito divulgado entre a populaça»...) é o «em cima da mesa», no sentido de um assunto estar — ou ser posto à — em discussão. Presumo que tal provenha do inglês do Reino Unido onde há o verbo «*to table*» (a palavra «*table*», mesa, está lá) com o sentido de submeter um assunto a debate. Parece que a classe política e uns que por aí andam têm o «em cima da mesa» em grande consideração... o que, aceita-se, não traz mal ao mundo (como, aliás, seria o caso se o arcebispo de Beja fosse ao Porto e dissesse que era o Napoleão![1]), mas aqui se quer assinalar, até para que não se repita até à eternidade.

E temos ainda o termo «tabu», que surge com frequência no teatro do quotodiano quando um ator da política não sabendo o que dizer — é difícil dizer coisas fundamentais e inovadoras... — difunde à puridade a ideia (que outros publicitarão como quem não quer a coisa, com ar recatado de virgem fora de prazo) de que tem algo de essencial a dizer, mas que ainda não chegou o grande momento. Há assim umas erupções deste quilate em certas épocas, nisso colaborando a comunicação social, o que por vezes é ridículo... porque, desfeito o tabu, dali não veio grande coisa. Mas, enfim, a vida tem disto!

[Desculpem mas tenho de interromper estas considerações, porque veio até esta mesa tipo *ikea* com tampo desdobrável, *motu proprio*, sem invitação, um livro que estou a reler; vou neste passo:

> «*Il ne pouvait une fois de plus qu'aboutir à la même conclusion : décidément, les femmes étaient meilleures que les hommes. Elles étaient plus caressantes, plus aimantes, plus compatissantes et plus douces ; moins portées à la violence, à l'égoïsme, à l'affirmation de soi, à la cruauté. Elles étaient en outre plus raisonnables, plus intelligentes et plus travailleuses.*
> *Au fond, se demandait Michel en observant les mouvements du soleil sur les rideaux, à quoi servaient les hommes ? Il est possible qu'à des époques antérieures, où les ours étaient nombreux, la virilité ait pu jouer un rôle spécifique et irremplaçable ; mais depuis quelques siècles, les hommes ne servaient visiblement à peu près plus à rien.*
> *Ils trompaient parfois leur ennui en faisant des parties de tennis, ce qui était un moindre mal ; mais parfois aussi ils estimaient utile de faire avancer l'histoire, c'est-à-dire essentiellement de provoquer des révolutions et des guerres. Outre les souffrances absurdes qu'elles provoquaient, les révolutions et les guerres détruisaient le meilleur du passé, obligeant à chaque fois à faire table rase pour rebâtir. Non inscrite dans le cours régulier d'une ascension progressive, l'évolution humaine acquérait ainsi un tour chaotique, déstructuré, irrégulier et violent. Tout cela les hommes (avec leur goût du risque et du jeu, leur vanité grotesque, leur irresponsabilité, leur violence foncière) en étaient directement et exclusivement responsables. Un monde composé de femmes serait à tous points de vue*

[1] *D'après* Mário-Henrique Leiria.

> *infiniment supérieur ; il évoluerait plus lentement, mais avec régularité, sans retours en arrière et sans remises en cause néfastes, vers un état de bonheur commun.*»

Michel Houellebecq — *Les particules élémentaires*, autor que veio a ser barbaramente assassinado a folhas tantas de *La carte et le territoire*].

Mas, retomando:

O «por acaso» também é aplicado à toa. Não há muito, uma senhora de uma revista cor-de-rosa — onde a elite nacional produz declarações fulcrais para o nosso país — afirmava que o «marido era por acaso pai dos meus (dela) filhos»... Bom, ela lá sabe!

Outros usos há que, segundo me parece — mas eu não sei tudo — estão a cair em desusos: o «é assim» e o «prontos» (gosto de ambos, como o Maomé da carne de porco!), embora ainda se oiçam aqui e ali.

Desafio: outra palavra muito apreciada e que, há alguns anos, transbordou do desporto para a área dos recursos humanos. Qualquer gestor (e há paletes deles!), ou candidato a, descreve-se como gostando de «desafios» e que a sua vida é feita de «desafios». Aliás, quem não afirma aceitar «desafios» é melhor tirar o cavalo da chuva e ir pregar para outra freguesia. Nada tenho a opor... e ainda que tivesse...; mas vejam lá se não abusam dos ditos e se variam o modo de expressão.

Mas não resisto ainda a referir de novo o que se vê amiúde nas legendas no cinema e na televisão: descarada falta de cuidado na tradução para português, com frases incompreensíveis e pejadas de erros de tradução, mormente quando se trata de incursões na área jurídica (aí é de ficar de *patas arriba*!). Só um exemplo: o inglês «*eventually*» traduzido para português por «eventualmente»... o que deu num caso uma tradução assim: «ele morreu eventualmente»... coisa que faz chorar as pedras da calçada e atirar pela janela fora a mobília de casa (pesiché incluído). Sinceramente, isto curiosamente parece-me estrategicamente mal posicionado!

luis.f.sabino@gmail.com

Hilaridade da precariedade

Jorge Madeira Mendes
Direcção-Geral da Tradução — Comissão Europeia

Em tempos não muito recuados, lia-se em cartazes de uma central sindical — e ouvia-se em intervenções dos seus dirigentes — a conspícua palavra «precaridade».

Ora, se bem que o *Dicionário da Língua Portuguesa* da Porto Editora (pelo menos na sua edição de 2009) a considere aceitável (ainda assim, com preponderância para «precariedade»), a verdade é que os grandes dicionários clássicos, nomeadamente os prestigiados *Aurélio* e *Houaiss*, ignoram por completo a grafia «precaridade». [No mínimo, hesitaria em tomar como referência a edição de 2009 do dicionário da Porto Editora, que consagra também «carenciado», neologismo desnecessário e aberrante (porque fruto de ignorância: com efeito, o adjetivo será «carecido» em caso de carência material, «carente» em caso de carência moral ou afetiva).]

Voltando ao universo da política portuguesa, apraz-me assinalar as honrosas exceções que aparentam fazer questão numa clara pronúncia de «precariedade».

Por outro lado, ocorre um curioso erro simétrico, provavelmente relacionado com a chamada «hipercorreção»: nos mesmos órgãos de informação que veiculam a «precaridade», ouve-se, paralelamente, falar em «complementariedade»
Existe uma mnemónica simplicíssima:

1) Os adjetivos terminados em *ar* ou em *are* (que são, por assim dizer, terminações simples) formam o correspondente substantivo acrescentando simplesmente o sufixo *idade*:

complementar	complementaridade
par	paridade
hílare	hilaridade;

2) Os adjetivos terminados em *ário* ou em *ório* (que são, comparativamente, terminações mais complexas) formam o correspondente substantivo acrescentando o sufixo *iedade* (mais complexo):

precário	precari**e**dade
solidário	solidari**e**dade
subsidiário	subsidiari**e**dade
notório	notori**e**dade.

Só falta convencermos os jornalistas e figuras públicas dos nossos meios de comunicação a renunciarem a «precaridades» e «complementariedades».

Jorge-Madeira.Mendes@ec.europa.eu

O termo precariado

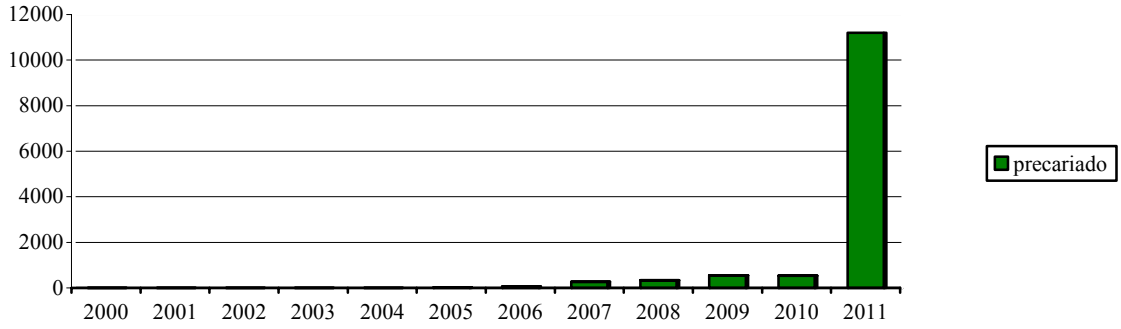
Marisa Gomes da Silva
Estagiária da Direcção-Geral da Tradução — Comissão Europeia

Segundo o texto de Luis González no n.º 124 de *puntoycoma*[(1)], as primeiras referências nas principais línguas europeias ao **precariado** datam do final do século XX[(2)]. Numa abordagem sociológica e semântica, pode dizer-se que este neologismo proveniente da sociologia e da política resulta da junção dos substantivos *precariedade* e *proletariado*.

Fazendo fé nos textos indexados pelo motor de pesquisa Google, a palavra precariado terá surgido pela primeira vez na língua portuguesa em 2000, e até 2006 é relativamente pouco mencionada. A partir de 2007 o seu uso aumenta, mas é apenas em 2011 que se verifica o maior crescimento, atingindo já mais de 11 200 referências, como se pode verificar no gráfico em baixo.

(1) *puntoycoma*. Índice numérico, http://ec.europa.eu/translation/bulletins/puntoycoma/numeros.html.

(2) Desde finais dos anos 90 o grupo dos *Precari Nati* elabora um discurso militante sobre *il precariato*. Em dezembro de 2006, a Fundação Friedrich Ebert, utilizou o termo *Prekariat* no seu estudo *Gesellschaft im Reformprozess*. No âmbito académico, o sociólogo Robert Castel analisou o fenómeno do *précariat* em *Les Métamorphoses de la question sociale: une chronique du salariat* (Fayard, 1995) e em obras posteriores. As referências ao termo foram frequentes nestes últimos anos em inglês, francês, alemão, italiano, português e espanhol, tanto na imprensa generalista como nos meios académicos e profissionais.



O ano de 2011 viu, em janeiro, a apresentação da canção dos Deolinda, «Parva que sou». A reação do público foi entusiasta e, nos dias que se seguiram, a letra da música foi citada frequentemente na arena política e social, tendo sido apelidada de «Hino ao Precariado»[3]. A Provedoria do Precariado[4] e os Precários Inflexíveis[5], juntamente com o Grupo Contra o Trabalho Precário, o Sindicato dos Precários[6] e o FERVE[7] são alguns dos movimentos frequentemente associados à contestação da precariedade. Recentemente foram responsáveis pela polémica à volta da pergunta 32 dos Censos 2011 (argumentando que os trabalhadores a recibos verdes não deveriam ser incluídos no mesmo grupo que os empresários); impulsionaram várias manifestações nacionais contra a precariedade, em conjunto com a CGTP Intersindical e participaram na reunião de assinaturas para levar a voto na Assembleia da República a Lei Contra a Precariedade. Presentes na rua, nos blogues, nas redes sociais e nos meios de comunicação contribuíram, assim, para um maior uso e reconhecimento da palavra em questão.

A formação do termo

O termo **precário**, do latim *precarius*, significa *obtido por meio de prece, que não é definitivo, portanto, provisório ou ainda escasso, insuficiente, difícil*. O *Dicionário Priberam da Língua Portuguesa*[8] reconhece a palavra não só como adjetivo mas também como substantivo masculino:

> *Indivíduo sem vínculo de trabalho permanente.*

Ao substantivo adiciona-se o sufixo nominal ***-ado***, do latim *-atu*, equivalente a *-ato* (anonimato, indigenato, baronato, tabelionato, artesanato). O sufixo transmite, neste caso em particular, um valor de estatuto ou condição e de coletividade (um conjunto — *eleitorado*; uma classe — *operariado*; um cargo — *inspetorado*). Morfologicamente, o sufixo *-ado* seleciona bases simples ou derivadas sufixadas em: *(d/s)or, **-ário**, -al, -ante, -ão,* etc.

Outros sufixos com o mesmo valor semântico são sobretudo *-ato,* mas também *-ia* e *-ura*. Uma vez que os sufixos *-ato* e *-ura* têm pouca utilização no português contemporâneo parece natural que, numa palavra tão recente, não sejam utilizados. A seleção de *-ado* em detrimento de *-ia* justificar-se-á, talvez, pela proximidade das palavras proletariado e precariado. Se precariado deriva de proletariado no que diz respeito ao seu significado sociológico e político, é aceitável que assim o seja também na sua morfologia.

---

[3] *Diário de Notícias* — «Deolinda vão editar o seu "hino" ao precariado», http://www.dn.pt/inicio/artes/interior.aspx?content_id=1775026&seccao=M%FAsica.
[4] Provedoria do Proletariado, http://provedordoprecariado.blogspot.com/.
[5] Precários Inflexíveis, http://www.precariosinflexiveis.org/.
[6] Sindicato dos Precários, http://www.facebook.com/group.php?gid=88867044575&ref=share.
[7] Fartos/as d'Estes Recibos Verdes — http://fartosdestesrecibosverdes.blogspot.com/.
[8] *Dicionário Priberam da Língua Portuguesa*, http://www.priberam.pt/dlpo/Default.aspx.

Chegamos, assim, à definição do dicionário Priberam de **precariado**:

> *Conjunto ou classe dos trabalhadores precários.*

Com base no *Novo Dicionário Eletrônico Aurélio* versão 5.11a, pode mesmo reforçar-se a ideia de uma formação idêntica à de proletariado:

> *proletariado / **precariado** — Substantivo masculino.*
> *1. A classe dos proletários / **precários**.*
> *2. Estado ou condição de proletário / **precário**.*
> *3. Camada social formada por indivíduos que se caracterizam por uma qualidade permanente de assalariados / **precários** e por modos de vida, atitudes e reações decorrentes de tal situação.*

Fica demonstrado que o termo precariado circula na esfera social e política, entre os meios de comunicação e a população em geral; é um termo reconhecido e utilizado internacionalmente, com correspondências em algumas das principais línguas europeias (ver quadro abaixo); pode já ser encontrado num dos dicionários mais utilizados da língua portuguesa; obedece às normas de formação de palavras e, finalmente, transmite um novo conceito, isto é, não existe outra palavra com o mesmo significado.

pt	fr	es	en	IATE
precariedade laboral	précarité de l'emploi	precariedad laboral	job insecurity	2224954
precário	travailleur précaire	precario	—	3538993
precariado	précariat	precariado	precariat	2231296

Com esta nota, que resume os principais argumentos e fontes utilizados para incluir o termo em português e noutras línguas na base de dados IATE, pretende-se sublinhar a importância de registar um termo utilizado frequentemente. Os dicionários gerais e vocabulários portugueses deveriam, talvez, tê-lo também em conta.

marisainesgomessilva@hotmail.com

A partícula sulf(o)-

Paulo Correia
Direcção-Geral da Tradução — Comissão Europeia

A base terminológica IATE e as memórias de tradução Euramis da Direção-Geral da Tradução da Comissão Europeia (DGT) contêm exemplos abundantes e variados de **terminologia química** em língua portuguesa[(1)]. No entanto, nem sempre a respetiva ortografia está correta — mesmo quando foi indicada por peritos nacionais.

[(1)] A este respeito, ver também o artigo «A partícula hidr(o)-» no n.º 36 d'«a folha», http://ec.europa.eu/translation/portuguese/magazine/documents/folha36_pt.pdf.

A verificação dos conteúdos da base IATE e das memórias Euramis[2] realizada em preparação para a aplicação do Acordo Ortográfico de 1990 (AO90) nas instituições da União Europeia permitiu localizar vários casos-tipo. Um desses casos é o da utilização como infixo ou sufixo de partículas iniciadas pela letra «s». Que fazer com o «s»? Dobrá-lo ou mantê-lo sozinho? A partícula «**sulf(o)-**» ilustra bem este problema.

Do girassol ao Eurosistema

O «**s**» no meio de palavras **a seguir a uma vogal** tem sempre de ser dobrado se se quiser manter o som «cê» que apresenta no início das palavras. Isso é aprendido logo na escola primária (*girassol*, etc.) e o uso impõe-no, com naturalidade, em palavras do vocabulário geral (*heterossexual*, *ecossistema*, etc.).

Em Portugal, em domínios técnicos em que a terminologia nos chega essencialmente por tradução do inglês (ou francês), parece, no entanto, haver uma certa tendência para a generalização de práticas ortográficas ao arrepio de qualquer reforma ou acordo ortográfico do português do século XX e do que já vai do XXI. Essa tendência vigente em certos meios técnicos — e, muitas vezes, reproduzida na tradução — tem sido a de aproximar (consciente ou inconscientemente) a mancha gráfica do português da mancha gráfica do original, respeitando, por exemplo, o número de «s» do original — numa convergência, certamente involuntária, com as regras do vizinho castelhano, onde o «s» não dobra.

> É, de alguma forma, também o caso do ***Eurosistema***, assim consagrado (em itálico!) na atual versão portuguesa consolidada do Tratado de Lisboa[3]. Na versão original do Tratado a ortografia utilizada havia sido Eurossistema[4]. De qualquer forma, sendo um nome próprio, pode admitir-se alguma maior liberalidade na (não) aplicação das regras ortográficas do português. No Brasil há um caso semelhante com o nome próprio **Mercosul**, outro caso notável da tal tendência «castelhanizante», que manteve o «s» simples de Mercosur.

A partícula «sulf(o)-» nas denominações de substâncias químicas

Muitos **termos químicos** em língua portuguesa incluídos na base IATE com referência a especialistas portugueses sofriam ou sofrem ainda dessa tendência «castelhanizante». O mesmo se passa em segmentos incluídos nas nossas memórias de tradução e em textos publicados no Jornal Oficial (JO). A ortografia dos termos com a partícula «sulf(o)-» presentes em fichas IATE da responsabilidade da Comissão Europeia foi revista e corrigida, dobrando o «s» quando no meio da palavra lhe corresponde o som «cê» e não o som «zê». As correções foram feitas com base em documentação especializada em língua portuguesa.

Na verificação da ortografia em fichas IATE de substâncias farmacêuticas que contêm a partícula «sulf(o)-» foram consultadas as listas de **denominações comuns internacionais (DCI) em língua portuguesa** do Infarmed — Autoridade Nacional do Medicamento e Produtos de Saúde[5] — e da

---

[2] Ver o artigo «A base IATE e as questões ortográficas» no n.º 36 d'«a folha» e a separata do n.º 35 d'«a folha», http://ec.europa.eu/translation/portuguese/magazine/documents/folha36_pt.pdf e http://ec.europa.eu/translation/portuguese/magazine/documents/folha35_vocabulario_pt.pdf.

[3] *Artigo 282.º*
1. O Banco Central Europeu e os bancos centrais nacionais constituem o Sistema Europeu de Bancos Centrais (adiante designado «SEBC»). O Banco Central Europeu e os bancos centrais nacionais dos Estados-Membros cuja moeda seja o euro, que constituem o *Eurosistema*, conduzem a política monetária da União.
http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:C:2010:083:FULL:PT:PDF.

[4] Ata de retificação do Tratado de Lisboa que altera o Tratado da União Europeia e o Tratado que institui a Comunidade Europeia, assinado em Lisboa em 13 de dezembro de 2007,
http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:C:2010:081:0001:0003:PT:PDF.

[5] Instituto Nacional da Farmácia e do Medicamento — Deliberação n.º 538/CA/2005: Denominação comum em Português (DCPt) da Denominação Comum Internacional (DCI) ou da Denominação comum (DC) de Substâncias Ativas, http://www.infarmed.pt/portal/pls/portal/docs/1/21083.PDF.

brasileira Anvisa — Agência Nacional de Vigilância Sanitária[6] — e posteriormente atendeu-se ao disposto no vade-mécum da Comissão da Farmacopeia Portuguesa[7]. A verificação do número CAS[8] (*Chemical Abstracts Service*) permitiu garantir a identificação de cada substância. Em anexo apresenta-se uma lista de fichas IATE da responsabilidade da Comissão.

Paulo.Correia@ec.europa.eu

IATE (inicial)	CAS	Infarmed/(Anvisa)	IATE (corrigida)	IATE
acediasulfona sódica	127-60-6	(acediassulfona sódica)	acediassulfona sódica	1885656
acesulfamo	33665-90-6	acessulfamo	acessulfamo	1893615
aldesulfona sódica	144-75-2	(aldessulfona sódica)	aldessulfona sódica	1886143
alusulfo	61115-28-4	(alussulfe)	alussulfe	1895337
aurotiosulfato sódico	10210-36-3	aurotiossulfato de sódio	aurotiossulfato de sódio	1884928
chaulmosulfona	473-32-5	(chaulmossulfona)	chaulmossulfona	1894289
diatimosulfona	5964-62-5	diatimossulfona	diatimossulfona	1895201
disulfirame	97-77-8	dissulfiram	dissulfiram	1884737
etasulfato de sódio	126-92-1	(etassulfato de sódio)	etassulfato de sódio	1885639
glucosulfamida	7007-76-3	(gluicossulfamida)	glucossulfamida	1895863
glucosulfona	554-18-7	(gluicossulfona)	glucossulfona	1892741
improsulfano	13425-98-4	(improssulfano)	improssulfano	1885857
manosulfano	7518-35-6	(manossulfano)	manossulfano	1896156
mesulfamida	122-89-4	(messulfamida)	messulfamida	1885516
mesulfene	135-58-0	(messulfeno)	messulfeno	1885880
picosulfato sódico	10040-45-6	picossulfato de sódio	picossulfato de sódio	1884869
piposulfano	2608-24-4	(pipossulfano)	pipossulfano	1893209
ritrosulfano	4148-16-7	(ritrossulfano)	ritrossulfano	1894097
salazosulfadimidina	2315-08-4	(salazossulfadimidina)	salazossulfadimidina	1893049
salazosulfamida	139-56-0	(salazossulfamida)	salazossulfamida	1885984
salazosulfatiazole	515-58-2	(salazossulfatiazol)	salazossulfatiazol	1894575
tiazosulfona	473-30-3	(tiazossulfona)	tiazossulfona	1894288
treosulfano	299-75-2	treossulfano	treossulfano	1893400
vanildisulfamida	119-85-7	(vanildissulfamida)	vanildissulfamida	1885427

(6) Agência Nacional de Vigilância Sanitária — *Denominações Comuns Brasileiras (DCB)*, http://www.anvisa.gov.br/medicamentos/dcb/index.htm e *Lista DCB 2007 – Consolidada*, http://www.anvisa.gov.br/medicamentos/dcb/lista_dcb_2007.pdf.

(7) Instituto Nacional da Farmácia — *Vademecum: Classificação Farmacoterapêutica de Medicamentos; Denominações Comuns das Substâncias Ativas de Medicamentos; Designações Normalizadas: Formas Farmacêuticas, Vias de Administração, Recipientes e Sistemas de Fecho*, http://www.infarmed.pt/portal/page/portal/INFARMED/PUBLICACOES/TEMATICOS/VADEMECUM/vademecum.pdf.

(8) Consultar Wikipédia — *Registo CAS*, http://pt.wikipedia.org/wiki/N%C3%BAmero_CAS.

Trad-Iure

Sofia Favila-Vieira; Constança da Camara Bobone
Ministério dos Negócios Estrangeiros

O Projeto *Trad-Iure* constitui uma das funcionalidades do novo Portal de Informação Legislativa — SIMPLEGIS —, cuja máxima é «*menos leis, mais acesso, melhor aplicação*» e surge como uma resposta direta à necessidade sentida pela **Presidência do Conselho de Ministros** de disponibilizar ao cidadão comum — pessoa ou empresa —, nacional e estrangeiro, uma ferramenta de apoio à compreensão, interpretação e/ou tradução da legislação nacional, capaz de contribuir para um melhor entendimento dos seus conteúdos e, assim, para o aumento da eficácia da mesma.

O XVII Governo Constitucional lançou no âmbito do SIMPLEX[(1)] um **programa de simplificação legislativa** intitulado SIMPLEGIS[(2)]. Fazia parte deste programa a criação de um portal de informação legislativa que, por um lado, integrasse numa única base de dados a informação atualmente disponibilizada através do *Diário da República eletrónico*[(3)] e do DIGESTO[(4)] e, por outro, disponibilizasse serviços e funcionalidades com utilidade para particulares e empresas.

A **Imprensa Nacional – Casa da Moeda** e os Serviços da **Presidência do Conselho de Ministros**, dos **Ministérios dos Negócios Estrangeiros,** da **Justiça** e da **Administração Interna**, bem como da **Procuradoria-Geral da República**, que integram o Grupo de Trabalho que concebeu, desenvolve e mantém o dicionário jurídico multilingue eletrónico ***Jurislingue***[(5)] foram convidados em maio de 2010 para desenvolver a funcionalidade daquele portal de informação legislativa, que foi definida como sendo um Glossário Jurídico de acesso fácil e imediato.

Como **Glossário/Tradutor Jurídico**, o *Trad-Iure* foi concebido para auxiliar todo aquele que, num dado momento, contacta com o Direito Interno, Comunitário e/ou Internacional, e necessita de material de apoio qualificado que contribua para a compreensão correta e, por conseguinte, o tratamento adequado da informação jurídica que lhe foi transmitida ou que quer transmitir.

Contém, entre outros, termos de **Direito Constitucional, Civil, Comercial, Penal, Administrativo e/ou Fiscal, Comunitário, Internacional e Internacional Privado**, bem como termos relativos à **Organização Judiciária e/ou Processual**.

Até à data, as línguas de trabalho do *Trad-Iure* são, para além do **português, o francês e o inglês**. A médio/longo prazo, pretende-se integrar, por ordem decrescente, o espanhol, o alemão, o neerlandês e o italiano. Sendo um projeto linguístico, atual e atualizado, incorporou no seu conteúdo as **alterações previstas pelo Acordo Ortográfico de 1990**, que deverá entrar oficialmente em vigor em janeiro de 2012.

O adiamento do lançamento oficial do portal de informação legislativa acima referido e, por consequência, do *Trad-Iure*, por motivos que se prendem com as mudanças governamentais ocorridas em junho deste ano, não afetou o trabalho desenvolvido ao abrigo do Projeto *Trad-Iure*. Assim,

---

(1) SIMPLEX, http://www.simplex.pt/index.asp.
(2) Governo — Programa de Simplificação Legislativa (SIMPLEGIS): Apresentação, http://www.portugal.gov.pt/pt/GC18/Documentos/PCM/simplegis.pdf;
SIMPLEGIS: Perguntas e Respostas, http://www.portugal.gov.pt/pt/GC18/Governo/Ministerios/PCM/MP/ProgramaseDossiers/Pages/20100510_MP_Prog_Simplegis.aspx.
(3) Imprensa Nacional-Casa da Moeda — *Diário da República eletrónico*, http://www.dre.pt.
(4) Imprensa Nacional-Casa da Moeda — Sistema Integrado para o Tratamento da Informação Jurídica (DIGESTO), http://digestoconvidados.dre.pt/digesto/.
(5) Gabinete de Documentação e Direito Comparado — *Jurislingue*, http://jurislingue.gddc.pt.

prevê-se que até ao final do ano este disponha de um acervo de **5000** termos nas línguas portuguesa, inglesa e francesa.

Os utilizadores deste glossário jurídico poderão efetuar pesquisas a partir de qualquer uma das três línguas indicadas, esclarecer dúvidas quanto à nova ortografia do português, estabelecer relações entre os termos, identificar a área do Direito a que pertencem e ainda sugerir a inserção e tradução de termos que considerem úteis e que não se encontrem no dicionário.

Tem-se aproveitado o atual compasso de espera para, com a experiência adquirida com o trabalho realizado até ao momento, introduzir melhorias que facilitem a consulta dos seus futuros utilizadores. A título de exemplo, refira-se a introdução de um critério de fiabilidade que permita estabelecer o grau de exatidão da tradução (de mera sugestão a tradução comprovadamente exata), de um campo de sinónimos, de um campo de observações e de uma fonte de referência para cada termo.

Por último, esperamos contribuir com o *Trad-Iure* para a correção, clareza e precisão da comunicação, através da consistência e harmonização da terminologia utilizada, bem como para um melhor conhecimento, no nosso país e no estrangeiro, de conceitos do Direito, designadamente português.

sofia.vieira@mne.pt
constança.bobone@mne.pt

Avaliação do sistema de tradução automática Moses no Departamento de Língua Portuguesa da DGT utilizando os *scripts Moses for Mere Mortals* — estudo de caso inglês-português

Maria José Machado; Hilário Leal Fontes
Direcção-Geral da Tradução — Comissão Europeia

[versão inglesa deste texto — http://ec.europa.eu/translation/portuguese/magazine/documents/folha37_moses_en.pdf]

Resumo
O Departamento de Língua Portuguesa (DLPT) da Direção-Geral da Tradução da Comissão Europeia testou o par linguístico inglês-português utilizando o sistema de tradução automática de código aberto Moses — instalado e operado com os *scripts Moses for Mere Mortals* — com vista a avaliar a sua utilidade para tradutores profissionais (para fins de publicação), mas também para os utilizadores gerais (para fins de compreensão).
Moses for Mere Mortals é uma aplicação de código aberto que constrói um protótipo de uma cadeia de tradução para o mundo real permitindo uma utilização bastante facilitada do sistema Moses, tornando-o assim acessível a um leque mais alargado de utilizadores.
Neste artigo apresenta-se os resultados obtidos com um *corpus* de 12,4 milhões de segmentos e dois *corpora* de teste compostos por um total de 136 documentos (cerca de 1 milhão de palavras). Foram treinados quatro motores com e sem otimização (*tuning)* e a tradução automática foi avaliada com os sistemas de avaliação automática BLEU e NIST e por avaliação humana.
Para a avaliação automática, foram selecionadas dez variantes de parâmetros para a avaliação dos oito motores. Para a avaliação humana, foram selecionadas cinco variantes de traduções Moses e foram recolhidos 16 500 juízos individuais para fins de tradução (1 a 5) e 16 500 para fins de compreensão (Sim/Não) de 11 avaliadores.
Os resultados são muito prometedores mesmo com estes motores básicos, que foram utilizados como uma ferramenta de tradução assistida por computador no fluxo de trabalho do DLPT para a produção de traduções durante cerca de um ano e meio.

1. Introdução

A Comissão Europeia apoia e incentiva o desenvolvimento em colaboração e a reutilização de aplicações de *software* de código aberto (F/OSS) financiadas por entidades públicas nas administrações públicas europeias através do seu *Open Source Observatory and Repository for European Public Administrations*[(1)] e atualmente num âmbito mais vasto com a sua Agenda Digital[(2)]. Os Programas-Quadro de Investigação e Desenvolvimento Tecnológico da União Europeia (UE) têm apoiado a investigação no domínio da tradução automática (TA) e, nomeadamente, o projeto EuroMatrix(Plus)[(3)], que disponibilizou o sistema Moses ao abrigo de uma licença LGPL[(4)]. Na última década, têm-se verificado grandes progressos na investigação neste domínio com o desenvolvimento da tradução automática estatística, um método de tradução automática que utiliza *corpora* para treinar o sistema.

A política de multilinguismo da UE produziu durante 50 anos um vasto corpo de textos multilingues de elevada qualidade e a Comissão Europeia tem investido em tradução automática e sido utilizadora durante os últimos 35 anos. A TA como uma ferramenta de tradução assistida por computador (TAC) pode contribuir significativamente para a Comissão cumprir a sua missão de tratar equitativamente todas as línguas na sua comunicação bidirecional com os cidadãos europeus e as empresas. A base Euramis é o sistema de memórias de tradução no qual são armazenados os alinhamentos dos textos traduzidos pelas instituições da UE. Estes *corpora*, com milhões de segmentos por língua, podem assim ser facilmente utilizados para treinar pares linguísticos com sistemas de tradução automática estatística.

2. Contexto

Os autores trabalham há mais de 20 anos no domínio da tradução e, durante cerca de 10 anos, utilizaram e contribuíram para o aperfeiçoamento do sistema ECMT baseado em regras. Por conseguinte, esta avaliação foi efetuada de uma forma pragmática e na perspetiva do tradutor, uma vez que os autores não têm formação formal no domínio do processamento da linguagem natural — os autores têm formação na área das línguas e da tradução e o autor dos *scripts* é um ex-tradutor licenciado em medicina.

3. Instalação e operação do sistema Moses com os scripts Moses for Mere Mortals

O sistema Moses foi instalado utilizando o conjunto de *scripts Moses for Mere Mortals* (MMM), desenvolvido por João Rosas em colaboração com os autores como avaliadores, e foi publicado sob uma licença GPL[(5)]. A versão do MMM utilizada foi a publicada no sítio Sourceforge do Moses, que instala a versão Moses de 14 de agosto de 2010. Neste estudo de caso utilizou-se um PC com 4 processadores e 8 GB de RAM.

O MMM gera um protótipo de cadeia de tradução formada por: Moses + IRSTLM + RandLM + MGIZA. Estes *scripts* não permitem fazer o treino com inclusão de informações linguísticas (*factored training)*. Os *scripts* MMM correm em Linux (distribuição Ubuntu) e automatizam as tarefas de instalação, criação de um conjunto representativo de ficheiros de teste, treino, tradução e avaliação automática.

---

(1) Open Source Observatory and Repository for European Public Administrations, http://www.osor.eu/.
(2) Comisssão Europeia — *Digital Agenda for Europe*, http://ec.europa.eu/information_society/digital-agenda/index_en.htm.
(3) EuroMatrix project: Statistical and hybrid machine translation between all European languages (2006-2009) http://www.euromatrix.net/; EuroMatrixPlus project (2009-2012): http://www.euromatrixplus.net/.
(4) Statistical Machine Translation: Software — *Moses*, http://www.statmt.org/moses/.
(5) mosesdecoder / scripts / moses-for-mere-mortals, https://github.com/moses-smt/mosesdecoder/tree/master/scripts/moses-for-mere-mortals; moses-for-mere-mortals, http://code.google.com/p/moses-for-mere-mortals/; Wikipédia — *Moses for Mere Mortals*, http://pt.wikipedia.org/wiki/Moses_for_Mere_Mortals.

Os principais objetivos do MMM são: 1) ajudar a construir um protótipo de uma cadeia de tradução para o mundo real; 2) guiar os primeiros passos dos utilizadores que estão a começar a utilizar o sistema Moses, disponibilizando-lhes um *Help-Tutorial* de fácil compreensão, bem como um guia de instalação rápida e uma demonstração; 3) treinar *corpora* de grandes dimensões; 4) traduzir documentos (por lotes); 5) permitir uma avaliação fácil e rápida do sistema Moses com as aplicações de avaliação automática BLEU e NIST (por lotes), tanto em relação a todo o documento como segmento a segmento (muito útil para uma avaliação humana rápida dos segmentos com melhor/pior pontuação BLEU/NIST); 6) integrar a tradução automática com as memórias de tradução.

O MMM é composto por seis *scripts*: ***create*** (compila o sistema Moses e os pacotes que este utiliza com um único comando); ***make-test-files*** (cria ficheiros de teste a partir do *corpus* bilingue de base); ***train*** (efetua todas as fases do treino a partir do *corpus* bilingue); ***translate*** (traduz um ou mais documentos com os parâmetros selecionados); ***score*** (faz a avaliação automática de um ou mais ficheiros); ***transfer-training-to-another-location*** (permite transferir todos os dados de um treino para outro local no mesmo computador ou para outro computador).

No MMM estão incluídas duas aplicações complementares (*add-ins*) em ambiente Windows — *Extract_TMX_Corpus* (ETC) e *Moses2TMX* — para completar a cadeia desde o documento original em formato Word até aos ficheiros TMX para importação pelo tradutor para a aplicação que utiliza as memórias de tradução.

O MMM também inclui uma lista das principais abreviaturas para português («*Nonbreaking_prefix file for the Portuguese language*»). Os *scripts* MMM permitem a adaptação dos principais parâmetros do sistema Moses (cerca de 80) a pares linguísticos e *corpora*/documentos específicos.

Neste artigo, apresenta-se um estudo de caso que demonstra como meros mortais podem utilizar o sistema Moses.

4. Dados para treino e otimização (tuning)

Para fins de treino, foi utilizado um *corpus* de 12,4 milhões de segmentos extraído da base de dados Euramis da DGT pela Unidade de Informática. O *corpus* contém todas as traduções da DGT de documentos da Comissão e toda a legislação e jurisprudência alinhadas e armazenadas até novembro de 2009.

Este *corpus* inglês-português contém 468,9 milhões de palavras (en+pt) e 12 468 232 segmentos bilingues, tendo sido limpo de carateres de controlo, de segmentos em que o segmento da língua de partida e da língua de chegada eram idênticos e de pares de segmentos com rácios de *tokens* superiores a 4:1. Não se procedeu à fusão de segmentos idênticos.

A componente portuguesa do *corpus* foi utilizada para o treino do modelo linguístico com os modelos linguísticos IRSTLM e RANDLM. Os *corpora* utilizados para otimização com 800 e 2000 segmentos eram compostos por extrações de segmentos provenientes de documentos de uma grande variedade de domínios/direções-gerais selecionados pela sua qualidade e não contidos no *corpus* de treino (CT).

5. Motores

Foram treinados quatro motores com diferentes parâmetros e que foram subsequentemente otimizados com os *corpora* de 800 ou 2000 segmentos (Quadro 1) com parâmetros por defeito, exceto quando indicado o contrário[(6)].

[(6)] Ver «*Default parameters*» no documento «*Help-Tutorial*» em https://github.com/moses-smt/mosesdecoder/tree/master/scripts/moses-for-mere-mortals/docs.

ID do motor	Modelo linguístico	n-gramas	Outros parâmetros alterados
E1	IRSTLM	7-gramas	*Tuning*: não
E1t	IRSTLM	7-gramas	*Tuning*: 800
E2	IRSTLM	7-gramas	*Tuning*: não; *Smoothing*: *improved Kneser Ney*
E2t	IRSTLM	7-gramas	*Tuning*: 800; *Smoothing*: *improved Kneser Ney*
E3	RANDLM	7-gramas	*Tuning*: não
E3t	RANDLM	7-gramas	*Tuning*: 2000
E4	RANDLM	9-gramas	*Tuning*: não; *MaxLen*=80
E4t	RANDLM	9-gramas	*Tuning*: 2000; *MaxLen*=80

Quadro 1. Motores treinados com parâmetros MMM por defeito, exceto quando indicado o contrário.

6. Corpora *de teste*

Testaram-se dois conjuntos de documentos (Quadro 2). O *corpus* de teste 1 continha 88 documentos que foram avaliados individualmente por 32 tradutores do DLPT quanto à sua utilidade para o trabalho de tradução. Esta avaliação foi efetuada durante um período de três meses durante o qual todos os tradutores puderam solicitar uma tradução Moses a utilizar em lugar do nosso sistema de tradução automática à base de regras então disponível (ECMT).

Por conseguinte, estes documentos não foram escolhidos em função de critérios específicos. Abrangem um amplo leque de direções-gerais/domínios (20) e a tradução Moses utilizada nestes documentos foi obtida com um motor anterior treinado com um *corpus* de 6,6 milhões de segmentos que continha documentos traduzidos na DGT num período de tempo mais curto e treinado com o modelo linguístico IRSTLM (*Witten-Bell smoothing*). O *corpus* de teste 2 era composto por 48 documentos selecionados entre documentos traduzidos por colegas que não eram utilizadores de TA na altura (11 tradutores).

O único outro critério utilizado na seleção foi abranger uma vasta gama de direções-gerais/domínios (19). Ambos os *corpora* de teste continham documentos representativos do nosso trabalho (tanto documentos legislativos como não legislativos), nomeadamente regulamentos, decisões, recomendações, comunicações, relatórios, Relatório Geral da Comissão, pareceres, documentos de trabalho dos serviços da Comissão, memorandos, programas, etc.

Conjuntos de documentos	N.º de páginas (internas)	N.º de palavras	N.º de segmentos	N.º médio de palavras por segmento	Segmentos com concordância a 100% com o CT (n.º)	Segmentos com concordância a 100% com o CT (%)
***corpus* de teste 1**	2 594	675 313	34 179	19,8	8 249	24,1%
***corpus* de teste 2**	1 250	349 026	17 234	20,2	3 635	21,1%
Total	3 844	1 024 339	51 413	19,9	11 884	23,1%

Quadro 2. Conjuntos de documentos utilizados para teste.

7. Testes com diferentes parâmetros de tradução

O *script* ***translate*** permite uma definição fácil de 17 parâmetros de tradução que podem ter um impacto significativo na qualidade dos resultados. Foram testadas várias combinações destes parâmetros com uma pequena amostra a fim de determinar se havia uma melhoria no desempenho do Moses. Após testes preliminares com diferentes combinações, foram selecionados 10 parâmetros para testes adicionais: *weight_t* (*wt*), *weight_l* (*wl*), *weight_d* (*wd*), *weight_w* (*wp*), *mbr, searchalgorithm* (*searchalg*), *cubepruningpoplimit* (*cubeprun*), *stack, maxphraselength* (*mpl*) e *distortionlimit* (Quadro 3).

Variantes	Parâmetros por defeito, exceto:									
	wp	*wd*	*wl*	*wt*	*searchalg*	*cubeprun*	*stack*	*mpl*	*mbr*	*distortion limit*
Var. A	-1,3	—	—	—	—	—	—	—	—	—
Var. 1	-1	0,5	0,9	1,5	1	2000	2000	30	1	7
Var. 1A	-1,3	0,5	0,9	1,5	1	2000	2000	30	1	7
Var. 1B	-1,6	0,5	0,9	1,5	1	2000	2000	30	1	7
Var. 1C	-2	0,5	0,9	1,5	1	2000	2000	30	1	7
Var. 1D	-2,5	0,5	0,9	1,5	1	2000	2000	30	1	7
Var. 2	-1	0,5	0,9	1,5	1	2000	2000	30	1	9
Var. 2A	-1,3	0,5	0,9	1,5	1	2000	2000	30	1	9
Var. 2B	-1,6	0,5	0,9	1,5	1	2000	2000	30	1	9

Quadro 3. Combinações de parâmetros de tradução testados com o *corpus* de teste 2.

8. Avaliação automática

Os dois *corpora* de teste foram traduzidos pelos respetivos tradutores e foi seguidamente efetuada uma avaliação automática sem eliminação dos segmentos com uma percentagem de concordância elevada no *corpus* de treino.

O *script* ***score*** foi utilizado para obter pontuações BLEU e NIST destes documentos, individual e globalmente, uma vez que para nós é importante ter uma ideia do desempenho do Moses com documentos de tipos muito diferentes. Por conseguinte, foram efetuadas traduções Moses dos documentos individuais que foram avaliadas automaticamente e posteriormente os ficheiros foram fundidos e novamente avaliados a fim de obter uma pontuação global.

A melhor pontuação BLEU foi obtida com o motor E2-Var.1 em ambos os *corpora* de teste. A melhor pontuação NIST foi obtida com o motor E4-Var.1 e E2-Def.

A variação nas pontuações chega a atingir um máximo de 9,44 pontos BLEU, consoante os parâmetros de treino e tradução utilizados. Os parâmetros por defeito produziram consistentemente pontuações BLEU mais baixas do que algumas das variantes testadas, da ordem de -2,25 pontos BLEU a -9,44 pontos BLEU em comparação com a variante com melhor desempenho (E2-Var.1).

A simples alteração do parâmetro «*word penalty*» de 0 (valor por defeito) para valores entre -0,5 e -1,5 produziu melhores pontuações BLEU, o que parece lógico uma vez que o inglês é uma língua mais sintética do que o português.

A alteração do parâmetro «*word penalty*» conjugada com outros parâmetros também produziu pontuações BLEU um pouco superiores.

O pior desempenho de todos os motores otimizados foi confirmado por avaliação humana, o mesmo acontecendo com os motores treinados com RANDLM, não só nesta avaliação como também noutras avaliações não estruturadas efetuadas em documentos individuais.

Os resultados obtidos com os métodos de avaliação automática BLEU e NIST divergem significativamente como se pode ver pelos resultados (com posição) apresentados no Quadro 4. A avaliação humana da amostra de 300 segmentos (e outras avaliações não estruturadas) corroboraram as pontuações BLEU a nível global e por documento. Utilizou-se também a pontuação linha a linha do *script* ***score*** para avaliar as traduções Moses em documentos selecionados e constatou-se que, a nível de segmento, as pontuações BLEU não são tão «fiáveis», mas mesmo assim ajudam-nos a detetar problemas mais rápida e facilmente.

MOTOR/ VARIANTE	MODELO LINGUÍSTICO	*corpus* de teste 1						*corpus* de teste 2					
		BLEU	NIST	Dif. BLEU ***	Dif. NIST ***	Posição BLEU	Posição NIST	BLEU	NIST	Dif. BLEU ***	Dif. NIST ***	Posição BLEU	Posição NIST
E1-Def **	IRSTLM-WB	49,5	11,689	-3,1	-0,006	11	7	46,28	10,806	-2,26	0,1062	17	6
E1-Var.1 *		52,1	11,698	-0,5	0,0028	3	5	48,5	10,727	-0,04	0,0271	2	9
E1-Var.A			-					48,06	10,616	-0,48	-0,083	5	12
E1t-Def	IRSTLM-WB-t800	47,1	11,133	-5,6	-0,562	14	14	44,16	10,314	-4,38	-0,386	24	21
E1t-Var.1								41,44	9,3611	-7,1	-1,339	28	29
E2-Def **	IRSTLM-IKN	50,4	11,793	-2,3	0,0981	8	2	47,07	10,888	-1,47	0,1887	14	1
E2-Var.1 *		52,6	11,695			1	6	48,54	10,7			1	10
E2-Var.A *		52,4	11,615	-0,2	-0,08	2	8	48,01	10,584	-0,53	-0,116	7	15
E2-Var.1B								46,56	10,317	-1,98	-0,383	16	20
E2t-Def *	IRSTLM-IKN-t800	48,7	11,32	-3,9	-0,375	12	12	45,54	10,432	-3	-0,268	21	18
E2-Var.1								40,62	9,1721	-7,92	-1,528	29	30
E3-Def **	RANDLM-7g	48,4	11,436	-4,2	-0,26	13	11	45	10,544	-3,54	-0,156	22	16
E3-Var.A			-					46,94	10,396	-1,73	-0,309	15	19
E3-Var.1 **		51,4	11,58	-1,2	-0,115	6	9	47,74	10,59	-0,93	-0,115	10	13
E3-Var.1A								47,17	10,47	-1,37	-0,23	13	17
E3t-Def	RANDLM-7g-t2000	46,4	10,963	-6,2	-0,733	15	16	43,4	10,135	-5,14	-0,565	26	25
E3t-Var1			-					43,78	9,8432	-4,76	-0,856	25	27
E4-Def **	RANDLM-9g	46,2	10,976	-6,5	-0,72	16	15	43,01	10,146	-5,53	-0,554	27	24
E4-Var.1 **		51,4	11,826	-1,2	0,1309	7	1	47,73	10,867	-0,81	0,1669	11	3
E4-Var.A **		49,7	11,554	-3	-0,141	9	10	46,23	10,648	-2,31	-0,051	18	11
E4-Var.1A								48,03	10,827	-0,51	0,1274	6	5
E4-Var.1B *		52	11,728	-0,6	0,033	4	4	48,25	10,751	-0,29	0,0514	3	8
E4-Var.1C								47,87	10,585	-0,67	-0,114	9	14
E4-Var.1D								46,03	10,233	-2,51	-0,467	19	23
E4-Var.2								47,67	10,874	-0,87	0,1741	12	2
E4-Var.2A								47,97	10,843	-0,57	0,1433	8	4
E4-Var.2B		51,9	11,739	-0,7	0,0434	5	3	48,24	10,774	-0,3	0,0747	4	7
E4t-Def **	RANDLM-9g-t800	43,2	10,289	-9,4	-1,406	17	17	40,18	9,4847	-8,36	-1,215	30	28
E4t-Var.A								44,59	9,9333	-3,95	-0,766	23	26
E4t-Var.1		49,7	11,206	-3	-0,49	10	13	46,03	10,277	-2,51	-0,423	20	22

* Variante com avaliação humana dos 300 segmentos por 11 avaliadores;
** Variante com avaliação humana preliminar de 125 segmentos com 13 variantes por um avaliador;
*** Diferença de pontuações BLEU/NIST em relação ao motor com melhor pontuação (E2-Var.1).

Quadro 4. Pontuações globais BLEU e NIST para os *corpora* de teste 1 e 2.

9. Avaliação humana em condições reais de trabalho

Foi efetuada uma avaliação em condições reais de trabalho do *corpus* de teste 1 na qual participaram 32 tradutores aos quais foi solicitado que dessem a sua opinião sobre a utilidade da tradução Moses para o seu trabalho, relativamente a cada documento. Foi solicitada uma avaliação Sim/Não e um comentário em texto livre. Apenas dois tradutores que não eram (na altura) utilizadores de TA consideraram a tradução Moses inútil para documentos específicos (BLEU com E2-Var.1: 37,81 e 35,39), embora tenha havido tradutores que consideraram a tradução Moses útil mesmo com pontuações BLEU igualmente baixas.

Como era de esperar, os documentos com um maior número de segmentos com concordância a 100% relativamente ao *corpus* de treino obtiveram melhores pontuações, o que confirma que o sistema Moses realmente «aprende» com os dados com que é alimentado. Contudo, alguns documentos com um número muito baixo de segmentos com concordância a 100% apresentaram também pontuações razoavelmente boas (BLEU ≥ 45).

Não são apresentados neste artigo resultados pormenorizados por documento. No entanto, no Quadro 5 são apresentados alguns números relativos às pontuações por documento obtidas com o motor com melhor desempenho (E2-Var.1) (que não foi o motor utilizado para produzir as traduções Moses usadas na tradução desses documentos).

BLEU	***corpus* de teste 1**	***corpus* de teste 2**
Pontuação global	52,63	48,54
Pontuação mais elevada	74,79	80,22
Pontuação mais baixa	30,83	28,98
Pontuações ≥ 50,00	56 documentos (1577 páginas)	16 documentos (328 páginas)
Pontuações 40,00-49,99	27 documentos (836 páginas)	17 documentos (506 páginas)
Pontuações < 40,00	5 documentos (181 páginas)	15 documentos (416 páginas)

Quadro 5. Gama de pontuações BLEU para os *corpora* de teste 1 e 2, por documento.

10. Avaliação humana de uma amostra de 300 segmentos

Foi efetuada uma avaliação estruturada de uma amostra de 300 segmentos extraída de ambos os *corpora* de teste utilizando o *script* ***make-test-files*** do MMM. Antes de extrair os segmentos, fez-se correr um *script* desenvolvido por Michael Jellinghaus (*filtersentences.perl*) a fim de eliminar dos *corpora* de teste os segmentos com uma concordância a 100% relativamente ao *corpus* de treino. Fez-se então correr o *script* ***make-test-files*** que dividiu o ficheiro do *corpus* de teste 1 em 80 setores e o do *corpus* de teste 2 em 40 setores a fim de extrair pseudoaleatoriamente três segmentos/setor. Foi extraído um número suplementar de segmentos a fim de eliminar os que não tinham conteúdo de tradução (referências do Jornal Oficial, números e títulos curtos), criando assim uma amostra de 300 segmentos (200 do *corpus* de teste 1 e 100 do *corpus* de teste 2).

Um dos autores efetuou uma avaliação preliminar de 13 variantes (entre os motores com melhores e piores pontuações BLEU (identificados com (*) e (**) no Quadro 4) com 125 segmentos desta amostra a fim de selecionar os mais úteis para avaliação por um maior número de colegas tradutores. Esta avaliação preliminar corroborou os resultados da avaliação BLEU.

Considerando que a pontuação BLEU é o método de avaliação automática mais utilizado e que as nossas avaliações humanas anteriores tinham definitivamente demonstrado que valores do parâmetro «*word penalty*» entre -0,5 e -1,5 produziam melhores resultados, decidiu-se «confiar» nas pontuações BLEU. Por conseguinte, foram selecionadas cinco variantes para avaliação por 11 tradutores: quatro das variantes com melhor pontuação dos motores não otimizados e a variante com melhor pontuação entre os motores otimizados e os restantes parâmetros por defeito. Os resultados são apresentados no Quadro 6.

Na avaliação dos 300 segmentos por 11 tradutores, o nosso objetivo era avaliar segmentos como aparecem no nosso trabalho quotidiano, ou seja, sem qualquer seleção/limitação dos segmentos por comprimento, complexidade, caráter técnico ou quaisquer outros critérios (média de 23 palavras/segmento). A metodologia desta avaliação baseou-se sobretudo nas avaliações efetuadas no âmbito do Projeto EuroMatrix(Plus) com adaptações ao nosso contexto, objetivos e recursos. Sabe-se por experiência própria como é difícil avaliar segmentos longos com duas ou mais orações com diferentes traduções/erros em diferentes traduções Moses, mas são esses os tipos de segmentos que se têm de traduzir no nosso trabalho quotidiano.

Como os documentos destes dois *corpora* de teste abrangiam domínios muito diferentes e muitos eram de caráter técnico, para além do original e das traduções Moses, foi incluída uma tradução de referência no quadro de avaliação fornecido aos avaliadores, uma vez que estes não eram especialistas em todos esses domínios. Visto que estes segmentos foram extraídos aleatoriamente dos dois *corpora* de teste com cerca de 70 000 segmentos, não foi fornecido contexto. Trata-se de uma limitação que não pode ser evitada neste caso.

Os segmentos foram avaliados em termos da sua aceitabilidade para fins de tradução e compreensão. A avaliação para fins de compreensão é apresentada apenas a título indicativo, considerando que não foi efetuada em condições de laboratório, uma vez que os avaliadores tinham ao seu dispor o texto original e a tradução de referência (e alguns tinham até traduzido alguns desses documentos), o que pode influenciar a sua avaliação. No conjunto, obtiveram-se 16 500 juízos individuais para fins de tradução (pontuações 1 a 5) e 16 500 para fins de compreensão (Sim/Não) para as cinco variantes Moses selecionadas.

Critérios de avaliação (fornecidos aos avaliadores)

A. Pontuação dos segmentos segundo a sua aceitabilidade para fins de tradução

Esta avaliação destina-se a avaliar a aceitabilidade da tradução Moses (uma tradução possível), mesmo se essa não for a escolha que um determinado tradutor teria feito na sua tradução. Várias variantes podem ter a mesma pontuação. O objetivo é classificar essas variantes numa escala de 1 a 5, sobretudo quanto ao nível de qualidade para fins de tradução:

1 — **Mau**: muitas alterações necessárias para uma tradução aceitável; não se poupa tempo algum;
2 — **Assim-assim**: bastantes alterações, mas poupa-se algum tempo;
3 — **Bom**: poucas alterações; poupa-se tempo;
4 — **Muito bom**: apenas pequenas alterações, poupa-se muito tempo;
5 — **Totalmente correto**: poderia ser utilizado sem qualquer alteração, apesar de o tradutor poder ainda fazer algumas alterações se fosse a sua tradução de um documento.

A pontuação 5 deve ser reservada para as traduções que poderiam ser utilizadas sem qualquer alteração, de um ponto de vista linguístico e de conteúdo.

B. Veredicto quanto à aceitabilidade de cada segmento para fins de compreensão (assimilação)

Avaliar se todo o significado do segmento pode ser entendido, mesmo que o segmento não esteja fluente/correto de um ponto de vista linguístico (1 — **Sim**; 0 — **Não**). Se a tradução contiver uma ou mais palavras na língua original que deveriam ser traduzidas, o veredicto deve ser «Não» (0) dado que, para efeitos de assimilação, tem de se ter em consideração que o utilizador pode não compreender uma única palavra da língua de partida (embora este não seja o caso nesta avaliação com en-pt).

11. Resultados

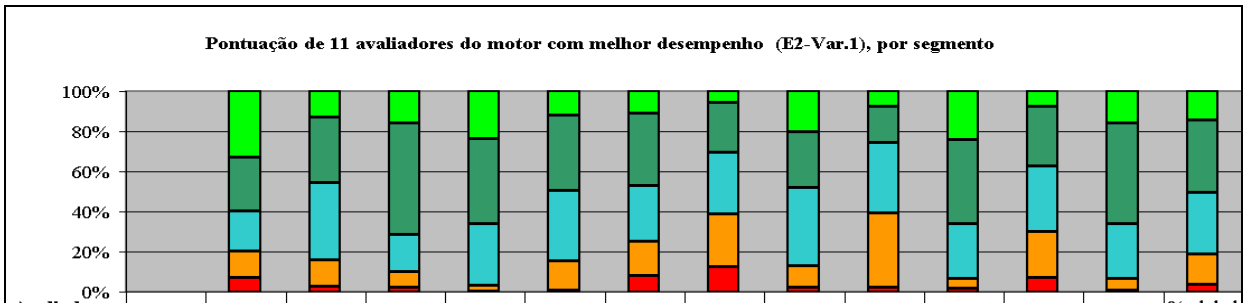
Os resultados globais são apresentados no Quadro 6 para os cinco motores/variantes avaliados, para fins de tradução e de compreensão, com uma percentagem global por avaliador e por variante.

No Quadro 7 são apresentados os resultados, por pontuação, do motor/variante com melhor desempenho (E2-Var.1), que teve uma percentagem global de 68,69 e 68,45 para fins de tradução e compreensão, respetivamente.

	E2-Var.1		E2-Var.A		E2t-Def.		E1-Var.1		E4-Var.1B	
Pontuação BLEU *corpus* de teste 1	52,63		52,43		48,73		52,09		52,00	
Pontuação BLEU *corpus* de teste 2	48,54		48,01		45,54		48,50		48,25	
Pontuação BLEU amostra de 300 segmentos	46,79		45,63		41,98		46,68		44,99	
	T (%)	**C (%)**	**T (%)**	**C (%)**	**T (%)**	**C (%)**	**T (%)**	**C (%)**	**T (%)**	**C (%)**
Média da avaliação humana (% de pontos) — global	68,69	68,45	68,19	66,94	63,62	59,48	67,30	66,58	65,65	62,73
Avaliador 1	68,07	69,67	67,47	69,67	62,93	58,33	67,00	67,67	65,53	63,33
Avaliador 2	75,27	69,33	78,80	66,00	70,93	58,33	79,20	65,67	72,93	60,00
Avaliador 3	76,20	74,67	75,80	72,67	72,33	66,33	75,07	71,00	73,13	66,33
Avaliador 4	69,27	52,00	68,60	49,67	65,80	45,33	67,60	49,00	67,27	48,00
Avaliador 5	65,13	73,33	61,67	71,00	54,80	61,67	59,33	70,67	62,53	69,33
Avaliador 6	57,13	48,00	55,53	44,67	52,60	37,00	55,33	45,67	53,87	42,67
Avaliador 7	70,80	68,67	70,13	65,00	63,47	51,67	70,13	67,33	67,53	62,00
Avaliador 8	58,60	87,00	61,20	89,33	57,00	87,33	58,67	90,33	58,87	89,33
Avaliador 9	78,20	76,33	75,40	75,33	70,20	67,00	74,33	79,67	70,93	69,00
Avaliador 10	61,80	68,33	59,93	66,00	56,00	59,67	57,53	62,67	55,20	58,67
Avaliador 11	75,13	61,00	75,60	67,00	73,73	61,67	76,13	62,67	74,40	61,33

T — tradução; C — compreensão

Quadro 6. Resultados da avaliação humana para fins de tradução e de compreensão da amostra de 300 segmentos extraída dos *corpora* de teste 1 e 2.



Avaliadores / Pontuação		1	2	3	4	5	6	7	8	9	10	11	% global segm.
Pontos: 5	5	39	49	71	37	34	18	61	24	73	24	49	14,48%
Pontos: 4	4	98	166	128	112	108	74	84	54	126	89	150	35,91%
Pontos: 3	3	116	56	92	106	83	92	117	105	82	97	82	30,94%
Pontos: 2	2	39	23	9	43	51	79	32	111	15	70	17	14,73%
Pontos: 1	1	8	6	0	2	24	37	6	6	4	20	2	3,48%

Quadro 7. Resultados, por tradutor e por pontuação, da avaliação para fins de tradução do motor/variante com melhor desempenho (E2-Var.1).

12. Conclusões

O sistema de tradução automática de código aberto Moses — instalado com os *scripts* MMM — demonstrou ser fiável e robusto, uma vez devidamente testados os *scripts*, traduzindo 81 000 páginas (cerca de 12 milhões de palavras — 1 milhão de segmentos) para este estudo de caso sem quaisquer problemas. A nossa principal preocupação é a qualidade, não a velocidade, mas também tomámos em consideração esse fator, o que explica o nosso interesse pelo modelo linguístico RANDLM. O modelo KenLM não foi testado neste estudo de caso.

Ficámos surpreendidos com os piores resultados obtidos de forma consistente com os motores otimizados, tal como demonstrado tanto pela avaliação automática como pela humana. Como a nossa abordagem é pragmática e obtivemos resultados satisfatórios com as variantes 1 e A dos parâmetros de tradução, passámos a utilizar a variante 1 no nosso fluxo de trabalho, uma vez que na avaliação se demonstrou ser ligeiramente melhor do que a variante A, em termos de qualidade, e é melhor em termos de velocidade no processo de tradução.

No que diz respeito à avaliação humana, e tendo em conta que testámos apenas motores de base, o nível de fluência e de precisão da terminologia foi surpreendentemente elevado e referido pela grande maioria dos tradutores/avaliadores. Em termos práticos, a utilização de TA (que já era elevada em 2009 com 85% dos tradutores portugueses a utilizar o sistema ECTM em pelo menos alguns documentos) aumentou com o sistema Moses e, em geral, verifica-se um elevado nível de satisfação entre os tradutores quanto à sua utilidade, conforme confirmado pelos resultados globais obtidos no que diz respeito à aceitabilidade para fins de tradução (superior a 60% com todos os cinco motores/variantes avaliados).

A coerência interavaliadores foi elevada em termos de classificação global (mas não de nível), uma vez que os motores/variantes considerados com melhor e pior desempenho foram, em geral, coerentes (nove avaliadores com a mesma opinião para cada um deles). A repartição por avaliador apresentada no Quadro 7 revela um elevado grau de variação entre pontuações, refletindo diferentes perceções dos tradutores quanto à utilidade para fins de tradução. Embora a avaliação para fins de compreensão seja apenas indicativa, correlaciona-se bem com a avaliação para fins de tradução.

Não procedemos a uma análise estatística aprofundada dos dados apresentados neste estudo de caso, nem estatísticas/análises a nível de segmento, visto que o nosso principal interesse é a utilidade do sistema Moses como uma ferramenta de tradução assistida por computador para utilização interativa no processo de tradução em combinação com as memórias de tradução. Parece haver uma boa correlação entre a avaliação automática e humana, no sentido em que as avaliações dos motores/variantes com melhor e pior desempenho em termos de pontuações BLEU foram corroboradas pela avaliação humana em geral.

Agradecimentos
Gostaríamos de agradecer ao nosso colega João Rosas, que desenvolveu e preparou os *scripts* que permitiram a realização deste estudo. Gostaríamos também de agradecer aos 32 tradutores das três unidades de tradução portuguesas que participaram ativamente nestas avaliações e que, em conjunto com os nossos outros colegas, nos dão continuamente informação sobre o desempenho do sistema Moses. Gostaríamos também de agradecer a Michael Jellinghaus, da DG TRAD do Parlamento Europeu, pelo *script filtersentences.perl*, que nos permitiu eliminar os segmentos com concordância a 100% quando da preparação da amostra de 300 segmentos para avaliação humana.
E agradecemos sobretudo aos membros da equipa de desenvolvimento do sistema Moses em todo o mundo que contribuíram para o desenvolvimento do sistema de código aberto Moses.

Referências
— Aziz, Wilker F.; Pardo, Thiago A. S.; Paraboni, Ivandré, «Fine-tuning in Portuguese-English Statistical Machine Translation». *Proceedings of the 2009 7th Brazilian Symposium in Information and Human Language Technology — STIL 2009*, São Carlos, 2009, http://portalsbc.sbc.org.br/download.php?paper=2816.
— Boyer, Vivian, «Human Evaluation of Machine Translation Quality: a Linguistic Oriented Approach», 2010.
— Callison-Burch, Chris; Koehn, Philipp; Monz, Christof; Peterson, Kay; Przybocki, Mark; Zaidan, Omar F., «Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation». *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR*, Uppsala, 2010, http://www.statmt.org/wmt10/pdf/wmt10-overview.pdf.
— Callison-Burch, Chris; Koehn, Philipp; Monz, Christof; Schroeder, Josh, «Findings of the 2009 Workshop on Statistical Machine Translation». *Proceedings of the 4th EACL Workshop on Statistical Machine Translation*, European Chapter of the Association for Computational Linguistics, Atenas, 2009, p. 1 a 28, http://homepages.inf.ed.ac.uk/pkoehn/publications/wmt09-overview.pdf.

— Callison-Burch, Chris; Fordyce, Cameron; Koehn, Philipp; Monz, Christof; Schroeder, Josh, «Further Meta-Evaluation of Machine Translation». *Proceedings of the Third Workshop on Statistical Machine Translation*, Association for Computational Linguistics, Columbus, 2008, p. 70 a 106, http://aclweb.org/anthology-new/W/W08/W08-0309.pdf.
— Callison-Burch, Chris; Fordyce, Cameron; Koehn, Philipp; Monz, Christof, Schroeder, Josh, «(Meta-) Evaluation of Machine Translation». *Proceedings of the Second Workshop on Statistical Machine Translation*, Association for Computational Linguistics, Praga, 2007, p. 136 a 158, http://www.statmt.org/wmt07/pdf/WMT18.pdf.
— Callison-Burch, Chris; Osborne, Miles; Koehn, Philipp, «Re-evaluating the Role of BLEU in Machine Translation Research». *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, 2006, p. 249 a 256, http://www.aclweb.org/anthology/E/E06/E06-1032.pdf.
— Caseli, Helena de Medeiros; Nunes, Israel Aono, «Tradução Automática Estatística baseada em Frases e Fatorada: Experimentos com os idiomas Português do Brasil e Inglês usando o toolkit Moses (NILC-TR-09-07)». *Série de Relatórios do Núcleo Interinstitucional de Linguística Computacional*, São Carlos, 2009, http://www2.dc.ufscar.br/~helenacaseli/pdf/2009/NILCTR-09-07.pdf.
— Eisele, Andreas; Federmann, Christian; Hodson, James, «Towards an effective toolkit for translators». *Proceedings of the 31rd Translating and the Computer Conference*, Association for Information Management, Londres, 2009, http://www.dfki.de/web/forschung/iwi/publikationen/renameFileForDownload?filename=TatC31.pdf&file_id=uploads_595.
— Jellinghaus, Michael; Poulis, Alexandros; Kolovratník, David, «Exodus — Exploring SMT for EU Institutions». *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR*, Uppsala, 2010, p.116 a 120, http://www.statmt.org/wmt10/pdf/WMT15.pdf.
— Koehn, Philip — *Statistical Machine Translation*. Cambridge University Press, janeiro 2010. ISBN-10: 0521874157.
— Koehn, Philipp; Birch, Alexandra; Steinberger,Ralf, «462 Machine Translation Systems for Europe». *Proceedings of the Machine Translation Summit XII*, International Association for Machine Translation (IAMT), Association for Machine Translation in the Americas (AMTA), Ontário, 2009, http://www.mt-archive.info/MTS-2009-Koehn-1.pdf.
— Koehn, Philipp; Schroeder, Josh; Osborne, Miles, «Edinburgh University System Description for the 2008 NIST Machine Translation Evaluation». *NIST Open Machine Translation 2009 Evaluation (MT09)* (collocated event with Machine Translation Summit XII), Ontário, 2009, http://homepages.inf.ed.ac.uk/pkoehn/publications/mteval08-report.pdf.
— Koehn, Philipp; Monz, Christof, «Manual and Automatic Evaluation of Machine Translation between European Languages». *Proceedings of the Workshop on Statistical Machine Translation* (Human Language Technology Conference/North American chapter of the Association for Computational Linguistics annual meeting), Association for Computational Linguistics, Nova Iorque, 2006, p. 102 a 121, http://www.aclweb.org/anthology-new/W/W06/W06-3114.pdf?CFID=68022703&CFTOKEN=25854326.
— Kos, Kamil; Bojar, Ondřej, «Evaluation of Machine Translation Metrics for Czech as the Target Language». *The Prague Bulletin of Mathematical Linguistics*, n.º 92, Dezembro 2009, p. 135 a 147, http://ufal.mff.cuni.cz/pbml/92/art-pbml92-kos-bojar.pdf.
— Nunes, Israel Aono; Caseli, Helena de Medeiros, «Primeiros Experimentos na Investigação e Avaliação da Tradução Automática Estatística Inglês-Português». *Anais do I TILic - Workshop de Iniciação Científica em Tecnologia da Informação e da Linguagem Humana* (collocated event with STIL 2009), 2009, São Carlos, p. 1 a 3.
— Specia, Lucia; Raj, Dhwaj; Turchi, Marcho, «Machine Translation Evaluation versus Quality Estimation». *Machine Translation*, SpringerLink, vol. 24, n.º 1, 2010, p. 39 a 50.

As memórias de tradução e a ortografia

Hilário Leal Fontes; Paulo Correia
Direcção-Geral da Tradução — Comissão Europeia

> «As instituições, órgãos e organismos da União Europeia decidiram aplicar, a partir de 1 de Janeiro de 2012, o Acordo Ortográfico da Língua Portuguesa de 1990. A partir dessa data, os textos publicados no *Jornal Oficial da União Europeia* serão redigidos segundo as regras da nova ortografia, admitindo-se um período inicial de coexistência das duas ortografias.»
>
> [Aviso a publicar durante o mês de dezembro de 2011 no Jornal Oficial (séries L e C)]

Há que pensar em como repercutir nas ferramentas ao dispor do tradutor português da Direção-Geral da Tradução (DGT) as mudanças decorrentes da aplicação do Acordo Ortográfico da Língua Portuguesa de 1990 (AO90) a partir de 1 de janeiro de 2012.

Todos os tradutores dispõem já de dicionários e vocabulários segundo o AO90 e de um corretor ortográfico também preparado para a nova ortografia. Os conteúdos IATE da responsabilidade da DGT estão a ser paulatinamente revistos e corrigidos. Restam as **memórias de tradução do serviço Euramis**[(1)], instrumento central do trabalho dos tradutores da DGT[(2)].

As memórias de tradução, que contêm cerca de 900 000 palavras diferentes, segundo a extração realizada em setembro de 2010, deverão permanecer com a antiga ortografia até dezembro de 2011, altura em que haveria duas possibilidades:

- ou manter as memórias tal como estão, aceitando que os tradutores devam proceder, durante muitos e muitos meses, à correção ortográfica permanente e repetitiva de palavras como *dire(c)tiva*, *prote(c)ção*, *a(c)tividades*, *a(c)ção*, *obje(c)tivo*, etc., que aparecem frequentemente nos segmentos provenientes das nossas memórias de tradução, até que se tenha constituído uma massa crítica de memórias criadas com textos pós-2012 que venha «sobrepor-se» às memórias de tradução pré-2012;
- ou proceder à alteração nas memórias da Comissão Europeia das palavras que mudam, evitando a correção ortográfica permanente e repetitiva de palavras frequentes pelos tradutores.

No departamento de língua portuguesa da DGT foi decidido adotar-se a última possibilidade, que parece ser a mais realista se quiser facilitar-se a introdução da nova ortografia sem afetar a qualidade e a produtividade. As alterações abrangerão todos os segmentos das memórias da DGT, mesmo os resultantes do alinhamento de legislação pré-2012[(3)].

Como tratar cerca de 900 000 palavras diferentes?

Os conversores ortográficos à nossa disposição são muito úteis, mas, tal como se pretendeu demonstrar no artigo «Conversores ortográficos e vocabulário das memórias de tradução»[(4)], requerem uma verificação *a priori* (AO45) e *a posteriori* (AO90) dos resultados. Não sendo possíveis essas verificações, num intervalo de tempo razoável, para a totalidade dos dados, foi decidido avançar-se por etapas, começando pelas palavras **mais frequentes** e identificando a lista das palavras a alterar.

---

(1) **Eur**opean **A**dvanced **M**ultilingual **I**nformation **S**ystem.

(2) Ficam excluídos os dicionários do sistema de reconhecimento vocal utilizado na DGT, que não podem ser alterados. Até à aquisição de um novo sistema já adaptado ao AO90, sugere-se a utilização em simultâneo do reconhecimento vocal e do corretor ortográfico com a ativação da opção «*Automatically use suggestions from the spelling checker*».

(3) Na reunião interinstitucional de 1 de março de 2011 dos serviço de tradução ficou decidido que se utilizará sempre o AO90 nas citações de textos redigidos antes da aplicação do Acordo, a fim de evitar duplas ortografias num mesmo texto.

(4) Equipa Linguística do Departamento de Língua Portuguesa — «Conversores ortográficos e vocabulário das memórias de tradução» in «a folha», n.º 36, Verão de 2011, http://ec.europa.eu/translation/portuguese/magazine/documents/folha36_pt.pdf.

Decidiu-se, ainda, alargar o trabalho à lista das palavras com **100 ou mais ocorrências** (**57 043 palavras diferentes**, face às 23 658 palavras diferentes das palavras com 500 ou mais ocorrências[(5)]). O tratamento de palavras com menor número de ocorrências conduziria a um volume de trabalho considerado incomportável[(6)].

A lista das palavras com 100 ou mais ocorrências foi **ordenada alfabeticamente** em formato «.txt» com o programa NotePad++ e **sujeita a conversão ortográfica** em formato «.doc», obtendo-se um ficheiro com cerca de 5 MB. A lista convertida foi estudada inicialmente para deteção e correção de conversões incorretas ou de omissões na conversão (por exemplo para certas palavras formadas por prefixação). Esse estudo permitiu, assim, determinar:

- **1.ª lista** — palavras que mudam com o AO90 (exemplo: *directiva* passa a *diretiva*);
- **2.ª lista** — gralhas e erros independentes do AO90 (exemplo: *ímans* em vez de *ímanes*);
- **3.ª lista** — variantes ortográficas passíveis de futura harmonização (exemplo: *sobresselente* e *sobressalente*).

Nestas três listas de palavras frequentes, identificaram-se as palavras que podem ser **emendadas automaticamente**. As listas parciais assim obtidas (com a antiga ortografia numa coluna e a nova ortografia noutra) foram, em seguida, enviadas ao gestor das memórias Euramis para ser efetuada a operação informática de substituição das formas ortográficas. A substituição será feita por etapas, abrangendo:

- **desde já** — gralhas e erros independentes do AO90 (permitirá testar o procedimento com uma lista limitada de palavras);
- **em finais de dezembro ou início de janeiro** — palavras que mudam com o AO90;
- **no futuro** — variantes ortográficas passíveis de harmonização.

N.B.: as palavras menos frequentes não são abrangidas nesta substituição; apesar de todo o cuidado posto na seleção das palavras sujeitas a alteração automática, não é de excluir que algumas homógrafas estrangeiras incluídas em segmentos portugueses possam ser tratadas indevidamente.

Caso haja interesse de outras instituições europeias em avançar também com esta alteração ortográfica automática controlada, o procedimento poderá ser facilmente alargado às memórias Euramis dessas instituições.

Ficam de fora palavras que devem ser **emendadas manualmente** com o editor *Sentence Manager* do Euramis. Foi, assim, preparada uma **4.ª lista** com estas palavras. É o caso de palavras que, em função do contexto, poderão ser ou não emendadas (exemplos: *eminente* usado no sentido de iminente ou de eminente; *sectorial* usado como setorial ou como a homógrafa inglesa ou espanhola *sectorial*).

Estas quatro listas podem ser consultadas no interior das instituições europeias no Wiki *Língua Portuguesa*[(7)].

O trabalho de verificação das palavras mais frequentemente utilizadas nas memórias de tradução Euramis em língua portuguesa permitiu igualmente melhorar outros **recursos** que poderão ajudar, a breve trecho, a aumentar a coerência dos textos produzidos na DGT:

---

(5) Equipa Linguística do Departamento de Língua Portuguesa — «Vocabulário das memórias de tradução», *in* «a folha», n.º 35, Primavera de 2011, http://ec.europa.eu/translation/portuguese/magazine/documents/folha35_vocabulario_pt.pdf.

(6) 83 427 palavras diferentes com 50 ou mais ocorrências; 206 475 palavras diferentes com 10 ou mais ocorrências; 319 117 palavras diferentes com 5 ou mais ocorrências.

(7) *Projeto Euramis PT COM AO90: Correções ortográficas das memórias de tradução Euramis* [acesso restrito às instituições europeias], http://intracomm.cec.eu-admin.net/wikis/display/languagept/Projeto+Euramis+PT+COM+AO90.

- lista de palavras corretas desconhecidas de corretores ortográficos[8];
- lista de erros ortográficos (utilizada na deteção de possíveis erros na base IATE).

Este trabalho terá também uma utilização imediata na adaptação do sistema de tradução automática *Moses* utilizado na DGT.

Hilario.Fontes@ec.europa.eu
Paulo.Correia@ec.europa.eu

Um CUSTOM.DIC à medida da DGT

Equipa linguística do departamento de língua portuguesa
Direcção-Geral da Tradução — Comissão Europeia

Os textos que saem da Direção-Geral da Tradução (DGT) em língua portuguesa devem estar ortograficamente corretos, em primeiro lugar, por se tratar de uma evidência e, em segundo lugar, porque, para muitos textos, não há qualquer «rede de segurança ortográfica» após a saída da DGT. É esse o caso de:

- textos publicados/enviados diretamente pela Comissão Europeia;
- textos revistos pelo Serviço das Publicações (OP) já em formato «.pdf» imediatamente antes da publicação (não há corretores ortográficos em língua portuguesa para os editores de «.pdf»);
- textos publicados na série C do Jornal Oficial em que apenas é feita uma verificação da concordância de conteúdos entre as várias línguas.

O **corretor ortográfico** é, pois, uma ferramenta de uso sempre obrigatório para os serviços de tradução e outros serviços das instituições europeias que redijam em língua portuguesa. O corretor será ainda mais importante nos primeiros tempos de aplicação do Acordo Ortográfico de 1990 (AO90), quando for necessário vencer antigos automatismos de escrita.

O corretor ortográfico disponibilizado com o processador de texto pode ser completado com **dicionários de utilizador** (CUSTOM.DIC). Esses dicionários individuais são alimentados sempre que, na janela de sugestões do corretor, se carrega no botão «*Add to Dictionary*». Com a aplicação do AO90 a partir de 1 de janeiro de 2012 **todos os dicionários de utilizador instalados nos nossos computadores devem ser revistos**. De outro modo palavras entretanto alteradas pelo AO90 continuariam a ser assinaladas como corretas sempre que o dicionário de utilizador estivesse ativado.

... à medida da DGT

Com a análise em curso das palavras presentes nas memórias de tradução Euramis verificou-se que há palavras frequentes nos nossos textos (100 ou mais ocorrências) que, embora ortograficamente corretas, não são reconhecidas pelo corretor ortográfico, mesmo quando estão ativados os vários dicionários temáticos disponíveis.

[8] Ver artigo «Um CUSTOM.DIC à medida da DGT», neste número d'«a folha».

Foi, assim, criado um ficheiro CUSTOM.DIC com essas palavras, às quais se juntou uma compilação de palavras recolhidas dos dicionários de utilizador postos à disposição da equipa linguística. **Todas as palavras foram adaptadas à nova ortografia**. Numa primeira fase, o ficheiro CUSTOM.DIC do Departamento de Língua Portuguesa não conterá nem nomes próprios nem estrangeirismos, a acrescentar após análise de possível harmonização de variantes ortográficas e/ou aportuguesamentos.

Como qualquer dicionário de utilizador, o CUSTOM.DIC do Departamento de Língua Portuguesa sofre de uma pequena limitação: **as entradas são sensíveis ao uso de maiúsculas e minúsculas e de flexões** (singular/plural e masculino/feminino). Deste modo, estando, por exemplo, *benzenossulfonato* registado no dicionário só com minúsculas, nem *Benzenossulfonato* nem *BENZENOSSULFONATO*, serão reconhecidos. Quanto às flexões, foram registadas as que ocorrem 100 ou mais vezes nas memórias de tradução, por exemplo, *aniónicos* é apenas registado no masculino plural, pois *aniónico*, *aniónica* e *aniónicas* são menos frequentes.

O **dicionário CUSTOM.DIC** do Departamento de Língua Portuguesa é disponibilizado como separata eletrónica deste número d'«a folha» em:
http://ec.europa.eu/translation/portuguese/magazine/documents/folha37_dicionario_pt.pdf.

Havendo interesse na sua utilização, é necessário salvaguardar e ativar o ficheiro CUSTOM.DIC em cada um dos computadores:

- guardar o ficheiro em «C:\Documents and Settings\(*login*)\Application Data\Microsoft\Proof», se necessário substituindo o anterior dicionário;
- ativar o dicionário do utilizador:
  - clicar em «Tools» + «Options» + «Spelling & Grammar» + «Custom Dictionaries»;
  - selecionar o último dicionário («CUSTOM.DIC (default)»), clicar em «Modify» e no campo «Languages» escolher «Portuguese (Portugal)»;
  - clicar em «OK».

DGT-PT-LINGUISTIC-TEAM@ec.europa.eu

Exoneração de responsabilidade: Os textos incluídos são da responsabilidade dos autores, não reflectindo necessariamente a opinião da Redacção nem das instituições europeias.
A Redacção é responsável pela linha editorial d'«a folha», cabendo-lhe decidir sobre a oportunidade de publicação dos artigos propostos.

Redacção: Paulo Correia (Comissão); Renato Correia (PE); Fernando Gouveia (TJ); Manuel Leal (Conselho da UE); Victor Macedo (CESE-CR); António Raúl Reis (Serviço das Publicações)
Grupo de apoio: Ana Luísa Faria (Conselho da EU); Hilário Leal Fontes (Comissão); Susana Gonçalves (Comissão); Ana Lorenzo Garrido (Comissão); Joana Seixas (CESE-CR)
Paginação: Susana Gonçalves (Comissão)
Envio de correspondência: dgt–folha@ec.europa.eu

Edição impressa: oficinas gráficas do Serviço de Infra-Estruturas e Logística — Bruxelas (Comissão)
Edição electrónica: sítio Web da Direcção-Geral da Tradução da Comissão Europeia no portal da União Europeia — http://ec.europa.eu/translation/portuguese/magazine

Os artigos contidos neste boletim podem ser reproduzidos mediante indicação da fonte e do autor.

«a folha» ISSN 1830-7809

ISSN 1830-780-9
9 771830 780004