Portuguese Vocabulary Profile: uma lista de vocabulário a aprendentes do PL2/PLE, baseada nos corpora de aprendentes e de livros de ensino

Shintaro Torigoe
Tokyo University of Foreign Studies / Gunma Prefectural Women's University

Abstract:
This paper reports the second pilot study of the *Portuguese Vocabulary Profile* (*PVP*) project, a Portuguese vocabulary list for learners in Japan based on the *Common European Framework of Reference for Languages*. Inspired by the *English Vocabulary Profile* (Capel, 2010, 2012), the *PVP* takes a learner-centric approach. For this study, the author modified the first pilot version which was constructed solely from learner corpora (Torigoe, 2016a) by comparing it with a word list based on a corpus of Portuguese textbooks published in Japan. The result is a broadened vocabulary for both the elementary and intermediate levels. The major improvement is that some intuitively basic words, including numbers, months of the year, foods, and facilities, which had been previously categorized as intermediate or advanced level words or which were missing from the first version due to their low frequency were correctly categorized as the elementary level words. However, the norm of word classification remains somewhat arbitrary given that the small size of both the input (learner corpora) and the comparative data (textbook corpus) does not allow for the use of statistical methods with less frequent words.

Keywords: word list for learners, CEFR, learner corpus, textbook corpus, Portuguese teaching in Japan.
Palavras-chave: lista de vocabulário a aprendentes, QECR, corpus de aprendentes, corpus de livros de ensino, ensino da língua portuguesa no Japão.

1. Introdução

No Japão, há mais de 150 mil falantes da língua portuguesa, que são predominantemente de origem brasileira os quais vivem e têm uma maior incidência populacional em determinadas regiões (Ministério de Assuntos Internos e Comunicações do Japão, 2009). Este facto qualifica o país como um observador associado da CPLP em 2014. Por consequência, a importância do português é cada vez mais reconhecida e a língua é cada vez mais ensinada tanto nas universidades como nas escolas secundárias japonesas. Segundo a pesquisa do Centro de Linguística da Universidade de Estudos Estrangeiros de Tóquio, 23 universidades japonesas têm curso de português.[1] Em adição, publicaram-se vários tipos de livros de ensino de português, particularmente desde o início da década de 2010 (Torigoe, 2015).

[1] http://www.tufs.ac.jp/common/fs/ilr/contents/lang/portuguese.html
Em adição, o autor identificou mais quatro universidades com o curso de português no ano lectivo de 2016.

No entanto, no Japão ainda não existe um critério de alvos de ensino e aprendizagem da língua portuguesa. Embora o Ministério de Ciência e Ensino do Japão (2009) organize os determinados *Courses of Study*, o currículo nacional do ensino do inglês, o qual define detalhadamente o que se ensina e se aprende em cada ano escolar, ainda não foi implementado para outras línguas estrangeiras, incluído português. Este facto representa e causa a desuniformidade nos livros didácticos de português. Relativamente à gramática, apesar de se poder identificar alguma diferença no objectivo final e ordem dos itens, os programas de estudo entre os livros didácticos são similares, normalmente os que são compostos conforme o tempo, o aspecto e o modo verbal. Contudo, relativamente ao vocabulário, verificam-se alguns casos que são influenciados pela área de estudo ou pela geração dos autores. Por exemplo, um livro de "vocabulário básico" contém nas primeiras páginas as palavras aparentemente não básicas, como *abismo*, *alfândega*, *besteira*, etc.

Tendo em conta o facto acima explicado, o autor reconhece a necessidade de criar para os aprendentes um catálogo de vocabulário mais objectivo possível. Este estudo, seguindo trabalho anterior do autor (Torigoe, 2016a), tem como objectivo propor uma segunda versão-piloto da lista de vocabulário, que o autor nomeia como *Portuguese Vocabulary Profile*. Na secção seguinte, o autor revê alguns exemplos de lista da vocabulário baseada nos dados objectivos.

1.1. De quantas e de que tipo de palavras precisam os aprendentes?

A quantidade e o tipo de palavras necessárias para os aprendentes constituem a questão principal do ensino de línguas estrangeiras. Segundo a revisão de Leiria (2006), as pesquisas em ensino e aprendizagem de vocabulário desenvolveram-se mais recentemente do que as de gramática. Para o efeito, existem actualmente algumas indicações no estudo de ensino de inglês baseado em corpora. Tono (2006, 2013) menciona que, do *British National Corpus*, os primeiros 100 vocábulos mais frequentes consistem em 67% das palavras totais do corpus, bem como os primeiros 1000 consistem em 87% e os primeiros 2000 consistem em 92%. Segundo o mesmo autor, mil vocábulos são equivalentes com o alvo do oitavo ano definido nos *Courses of Study* do Japão, enquanto 2000 vocábulos com o do décimo ano[2].

[2] No Japão, o ensino da língua estrangeira (inglesa) começa a partir do sétimo ano. Segundo a avaliação de Tono (2013), o alvo do oitavo ano é equivalente com o A2 do *QECR* , enquanto o alvo do décimo ano não alcança ao nível B1 do *QECR*.

1.2. Listas de Vocabulário

Influenciado pela estatística acima apresentada da autoria de Tono (2006, 2013), há vários livros de aprendizagem do vocabulário inglês seleccionado da informação de corpus, especialmente no Japão (e.g. Ishikawa, 2006; Tono, 2010). Na língua portuguesa, também há alguns exemplos similares. O primeiro exemplo é o *Léxico Multifuncional Computorizado do Português Contemporâneo* (*LMCPC*)[3], a lista de vocabulário baseada no *Corpus de Referências do Português Contemporâneo* (*CRPC*)[4]. O outro exemplo é *A Frequency Dictionary of Portuguese* (Davis & Preto-Bay 2008), o qual é um dicionário bilingue de português e inglês seleccionado do *Corpus do Português*[5]. Podemos ainda fazer uso de um outro exemplo no Japão. Aires & Iyanaga, (2012) é a lista de verbos selecionados da *LMCPC*, a qual já é aproveitada nas aulas da Universidade de Estudos Estrangeiros de Quioto.

1.3. Listas de vocabulário que se refletem do *QECR* (*CEFR*)

Como pudemos verificar previamente, quer em língua inglesa quer nas outras línguas (incluindo a língua portuguesa), já existem vários materiais de aprendizagem de vocabulário baseados em corpus. Contudo, uma dúvida permanece: as palavras usadas frequentemente por falantes nativos terão sempre correspondência com as palavras importantes para aprendentes?

O autor toma a posição oposta, baseando-se na ideologia desenvolvida e apresentada no *Quadro Europeu Comum de Referências* (*QECR*, *CEFR*; Council of Europe, 2001), o qual se divulga não só nos países europeus, mas também na fora da União Europeia, em regiões como Leste e Sudeste Asiático (Tono, 2013). O *QECR* estabelece o alvo de aprendizagem conforme o comportamento, que não é o que os professores têm de ensinar segundo a gramática tradicional mas o que os aprendentes de certo nível actualmente controlam. Relativamente ao ensino de vocabulário, igualmente, enquanto os materiais baseados no normativismo ou meramente na intuição de falantes nativos que procuram que tipo de palavras se deve ensinar, os materiais baseados no *QECR* procuram identificar que tipo de palavras os aprendentes sabem controlar.

[3] http://www.clul.ul.pt/sectores/linguistica_de_corpus/lmcpc/
[4] http://www.clul.ul.pt/pt/investigacao/183-reference-corpus-of-contemporary-portuguese-crpc
[5] http://www.corpusdoportugues.org/

Nº 3 – 09/ 2017 | 387-400 | **https://doi.org/ 10.26334/2183-9077/rapln3ano2017a20**

Revista da Associação Portuguesa de Linguística

No ensino da língua inglesa, o presente autor já tem confirmados dois exemplos de lista de vocabulário baseada no princípio do *QECR*, embora, ambos não adoptem os dados de aprendentes como dados primários, mas como dados comparativos.

O primeiro exemplo é o *English Vocabulary Profile* (*EVP*)[6] da autoria da Cambridge University Press dirigido por Annette Capel (2010, 2012). O *EVP* foi inicialmente selecionado das 5000 ou 6000 palavras "importantes" contidas no *Cambridge Learner Dictionary*, no qual foi baseado no corpus de falantes nativos (*Cambridge International Corpus*). Depois, estes primeiros dados foram modificados através da comparação com o *Cambridge Learner Corpus*, cujos subcorpora são anotados e divididos segundo os níveis do *QECR*, e também com corpora de livros didácticos. O destaque deste inventário de vocabulário é a marcação dos níveis do *QECR*, não apenas relativamente aos lemas mas também aplicada aos significados de palavra. No início do projecto, o *EVP* teve apenas os inventários do nível A1 ao B2 (Capel, 2010), mas ampliou-se até o nível C2 com a adição de dados com as colaborações de outros institutos (Capel, 2012).

O outro exemplo é a *CEFR-J Wordlist* de Tono (2013). Diferente do *EVP*, a fonte de vocabulário desta lista consiste em um corpus dos livros didácticos da língua inglesa publicados nos países do Leste da Ásia como a China, a Coreia do Sul e Taiwan. Numa primeira etapa, a equipa de pesquisa anotou os níveis do *QECR* nos livros de fonte e dividiu-os nos subcorpora pelos níveis. Numa segunda etapa, a equipa da pesquisa identificou as palavras estatisticamente significantes em cada nível. Tono pretendeu criar uma lista como o *EVP* orientada aos aprendentes do Japão, em que, segundo a pesquisa de Negishi, Takada & Tono (2012; referida em Tono, 2013), os aprendentes do nível iniciante e elementar dominam 80% de todos os aprendentes nacionais. Por esta razão, estabeleceu o nível Pre-A1 enquanto aboliu o C1 e C2.

Tono admite também a diferença entre o *EVP* e a *CEFR-J Wordlist* devido às características diferentes das fontes: enquanto o *EVP* é baseado no vocabulário produtivo, *a CEFR-J Wordlist* é baseado no vocabulário receptivo. Mas Tono menciona que o uso de ambas as listas permitirá a avaliação multiperspectiva de palavras. Além disso, actualmente Tono iniciou o projecto chamado *CEFR-J 27*, que procura criar uma lista de vocabulário nas 27 línguas, incluído português, que são ensinadas na Universidade de Estudos Estrangeiros de Tóquio.

[6] http://www.englishprofile.org/wordlists

2. Objectivo do estudo

O objectivo deste projecto é propor o *Portuguese Vocabulary Profile*, a versão portuguesa do *English Vocabulary Profile*, que é baseada não só nos dados de falantes nativos nem na intuição deles, mas também nos dados de aprendentes e materiais didácticos. Por conseguinte, este artigo tem como objectivo adicional refinar uma lista criada meramente dos dados de aprendentes comparando-os com os outros dados. Através do *PVP*, o autor pretende melhorar a qualidade dos livros didácticos e do ensino do PLE/PL2 no Japão.

3. Metodologia

A lista de vocabulário é criada através dos três procedimentos decorrentes: o primeiro é criação de uma lista de vocabulário de aprendentes; o segundo é a criação de uma lista de livros didácticos como os dados comparativos; o terceiro é comparar semiestatisticamente as duas listas e formar uma lista final. Diferente do *EVP,* que adoptou as entradas de dicionário, esta lista adopta os corpora de aprendentes como os dados primários de forma a tornar-se mais centrada em aprendentes. Por outro lado, utiliza os dados de vocabulário receptivo de livros didácticos como os dados comparativos já que, como mencionarei na próxima secção da minha investigação, adoptar apemas os dados produtivos causa a avaliação inadequada das palavras "fundamentais" mas menos usadas.

3.1. Lista de Vocabulário de Aprendentes

A lista de aprendentes (a Lista L2) já foi criada em Torigoe (2016a). Nesta secção, mostra-se o procedimento da criação da lista. A fonte da lista L2 é os *Corpora de PLE*[7]e o *Corpus de PEAPL2*[8]. Os dois corpora são os de composição por estudantes universitários obtidos através do mesmo procedimento. Os informantes são classificados nos níveis do *QECR* conforme as suas respostas de questionário. Seguidamente, mostra-se na Tabela 1 o detalhe estatístico dos corpora.

[7] http://www.clul.ul.pt/pt/recursos/314-corpora-of-ple
[8] http://www.uc.pt/fluc/rcpl2

	Corpora de PLE	*Corpus de PEALP2*
palavras totais	70.500	120.000
nacionalidades dos informantes	27	39
nº de subcorpora	471	546

Tabela 1: Perfis dos corpora de aprendentes

Dos dois corpora de aprendentes, foram identificados e extraídos 4857 lemas, incluído abreviações e palavras estrangeiras. Classificado por frequência, os primeiros 500 vocábulos mais frequentes dominam 90% das palavras totais dos corpora, bem como a frequência do 500º vocábulo é 32. Igualmente, os primeiros 1000 vocábulos dominam 95% do total e a frequência do 1000º vocábulo é 9, enquanto a frequência da 1493º vocábulo se torna menos de 5. Por estas razões, os pontos de divisão dos níveis foram colocados em cerca do 500º, a 1000º e a 1500º vocábulo.

No respeitante à adição da frequência, a metodologia estatística foi adoptada para marcar os níveis do *QECR* nos vocábulos. Em primeiro lugar, os 500 vocábulos mais frequentes no subcorpus do nível A1-A2[9] são automaticamente classificados no nível elementar, devido à alta correlação com os primeiros 500 do corpus total (r=0.96, p<0.01, pelo coeficiente de Pearson). Em segundo lugar, o exame de chi-quadrado foi utlizado; dos 1000 vocábulos seguintes, os vocábulos significativamente mais usados pelos aprendentes do nível A1-A2 do que os do nível B1-B2 foram adicionados no nível elementar (A1-A2). Igualmente, os vocábulos significativamente mais usados pelos aprendentes do B1-B2 do que os do Nível C1-C2 foram classificados no nível intermédio (B1-B2). O resto dos vocábulos com a frequência de mais de cinco foram classificados no nível avançado (C1-C2).

A adquirida lista de L2 é baseada exclusivamente na informação de frequência e estatística das palavras usadas nos corpora, portanto não contém ou estima inadequadamente as palavras com pouca utilização. Por exemplo, os vocabulários de número, tempo e cor que parecem intuitivamente básicos como *doze*, *dezembro*, *ontem*, *negro*, são classificadas como pertencentes ao nível intermédio ou até avançado; por outro lado, palavras do vocabulário brasileiro como *ônibus*, *trem*, *celular*, *gol*, *ruim*, ou até *você* aparecem poucas vezes no corpus e ficam fora da lista. Aliás, o tamanho do vocabulário em cada nível, que é cerca de 500

[9] Apliquei três divisões dos níveis (A1-A2, B1-B2 e C1-C2) conforme os *Corpora de PLE*, enquanto o *PEAPL2* adopta seis divisões.

palavras, é mais pequeno do que os do *EVP* e da *CEFR-J Wordlist* (veja na Tabela 2, secção 4.1).

Por estes pontos aqui debatidos, defendemos que a lista deveria ser modificada com dados de outro tipo como apresentarei na secção seguinte.

3.2. Lista de Vocabulário de Livros Didácticos

Nesta secção mostra-se o procedimento de criar a lista de vocabulário de livros didácticos (a Lista Tx). Para criar a lista, o autor construiu o corpus de livros didácticos (Torigoe, 2016b). A fonte do corpus é os livros de ensino de língua portuguesa publicados no Japão. Enquanto nesse país são actualmente disponíveis diversos tipos de material, adoptaram-se os livros "integrados", os quais consistem em diálogos modelos, explicações gramaticais e exercícios. Em dezembro de 2016, o autor confirma 11 livros deste tipo e já criou os subcorpora dos seis livros para iniciantes[10]. O corpus foi anotado com o *TreeTagger*[11], analizado com o concordancer *AntConc* (ver. 3.2.4w)[12] e finalmente arrumado com o *Microsoft Excel* para ser organizado como a lista de vocabulário.

O conteúdo da adquirida Lista Tx foi estatísticamente diferente da lista de vocabulário de falantes nativos. Comparado com o *LMCPC*, a Lista Tx contém mais vocabulários de vida quotidiana como, *morar*, *almoçar*, *passear*, *cansar*, *barato*, *aeroporto*, *descansar*, *supermercado*, *dançar*, *bicicleta*, *namorado*, *experimentar*, *biblioteca*, *aniversário*, etc. Por outro lado, faltam certas palavras frequentes no corpus de falantes nativos, por exemplo, as palavras sociopolíticas como *político*, *público*, *sociedade*, *internacional*, *caso*, *direito*, *economia*, *relação*, *social*, *força*, *condição*, *desenvolvimento*, *atividade*, *população*, *votar*; certos advérbios e preposições como, *apenas*, *durante*, *contra*, *quase*, *através*, *seguinte*, *conjunto*; e também certos verbos como *existir*, *tornar*, *viver*, *criar*, *realizar*, *utilizar*, que se podem substituir por verbos polissémicos.

Apesar da diferença relativa aos dados de falantes nativos, a Lista Tx contém vocabulários fundamentais que a Lista L2 avalia inadequadamente devido à frequência. Por este motivo, na próxima secção, utilizarei a Lista Tx como os dados comparativos à Lista L2 para organizar a lista final.

[10] Os pontos finais de cada livro são diferentes.
[11] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
[12] http://www.laurenceanthony.net/software.html

3.3. A lista final

Na etapa final desta investigação, o autor obteve a lista final através da comparação estatística entre a Lista L2 e a Lista Tx. Primeiro, eliminou-se da Lista L2 alguns jargões, nomes de local e palavras estrangeiras como *ERASMUS*, *Pusan*, *Alcalá*, *Manchester*, *new*, etc. Segundo, os vocábulos que não foram classificados no nível elementar da Lista L2 mas utilizados em mais de quatro livros, mais da metade do total, foram adicionados ao nível A (*segunda-feira*, *janeiro*, *amanhã*, *ônibus*, *celular*, *ruim*, etc.). Terceiro, os vocábulos que não satisfazem a condição do nível A mas foram usados mais de cinco vezes pelo aprendentes e usados por mais de dois livros foram classificados como pertencentes ao nível A+. Quarto, certos vocábulos do nível intermédio e avançado da Lista L2, os quais foram usados menos de cinco vezes pelos aprendentes do nível A1-A2, ou os quais não foram usados em nenhum dos livros didácticos, mantiveram-se classificadas como pertencentes ao nível B. Adicionalmente, os vocábulos que não aparecem na Lista L2 mas aparecem apenas uma vez na Lista Tx também foram classificadas como fazendo parte do nível B. Finalmente, como a *CEFR-J Wordlist*, não se estabeleceu o nível C.

4. Resultados

4.1. Observação quantitativa

Devido à limitação de espaço, a lista final não pode ser mostrada neste artigo mas na página da Internet elaborada para o efeito[13]. Nesta secção observa-se o detalhe estatístico da lista final através da comparação com a primeira Lista L2, com o *EVP* e a *CEFR-J Wordlist* (Tabela 2). O total do vocabulário do nível A e A+ é 1445, isto corresponde a cerca do triplo do nível elementar da Lista L2 e é quase igual com o mesmo nível do *EVP*, embora um pouco menos do que o da *CEFR-J Wordlist*. O total do nível B, que é 1487, também tornou-se quase o triplo do nível intermédio da Lista L2, no entanto, é menos da metade do mesmo nível do *EVP* e a *CEFR-J Wordlist*.

[13] www.tufs.ac.jp/ts2/society/pvp

	Lista final	Lista L2	*EVP*	*CEFR-J Wordlist*
A1	A: 911	531	601	1000
A2	A+: 534		925	1000
B1	1487	417	1429	2000
B2			1711	2000
C1-C2	n/a	472	2300	n/a

Tabela 2: Números dos vocábulos de cada corpora

4.2. Observação qualitativa

Esta secção, focaliza-se no vocabulário qualitativamente, observando a mudança dos níveis entre as primeiras listas (L2 e Tx) e a lista final.

Da Lista L2, 145 vocábulos do nível intermédio e 79 do nível avançado tornaram-se do nível A da lista final. De vários tipos, são destacáveis os vocabulários de comida e refeição, de estabelecimento e infraestratura, de emprego e trabalho, de relação pessoal e de clima (Tabela 3).

Tipo	palavras
comida e refeição	peixe, prato, fruta, chocolate, água, massa, sopa, delicioso, fome, sobremesa, leite, chá, arroz, batata
estabelecimento e infraestratura	aeroporto, supermercado, igreja, piscina, construir, museu, prédio, hospital, ponte, biblioteca, mercado, hotel, universitário
emprego e trabalho	médico, empresa, emprego, serviço, escritor, profissão, tarefa, chefe
relação pessoal	relação, grupo combinar, abraço, mandar, pedir, receber, comunicação, conversar, acordo, namorada, convidar, aceitar, perguntar, obrigar, etc.
clima	vento, chover, clima, temperatura, céu, chuva

Tabela 3: Vocabulários do nível A que foram dos níveis intermédio e avançado

A seguir, verificaremos as palavras que não apareceram na Lista L2 mas contudo foram classificadas no nível A pelo uso na ampla variedade (mais de quatro) dos livros. A maior parte pertence aos vocabulários de número, de calendário e de cor. Por outro lado, como se verifica nos vocábulos adicionados do nível intermédio e avançado, são também destacáveis os vocabulários de comida e refeição, de emprego e trabalho, de relação pessoal (Tabela 4).

tipo	palavras
comida e refeição	açúcar, churrasco, faca, feijoada, frango, salada, caipirinha, cebola, colher, gostoso, queijo, sal, uva, etc.
emprego e trabalho	reunião, vender, advogado, enfermeiro, funcionário, garçom, geladeira, guia, intérprete, etc.
relação pessoal	desculpar, senhor, esposo, compromisso, convite, dom, neto, obrigado, etc.

Tabela 4: Vocabulários do nível A que não foram usados pelos aprendentes

Por outro lado, há 27 vocábulos no nível A que não aparecem nos livros didácticos. A maior parte destes vocábulos pertencem ao vocabulário do português europeu e de nomes de locais (Tabela 5).

tipo	palavras
português europeu	autocarro, comboio, desporto, baixa, metro, porquê, facto
locais	Coimbra, Madrid, Berlim, Figueira, Turquia, Algarve

Tabela 5: Vocabulários do nível A que não apareceram nos livros didácticos

Além do nível A, foram adicionados mais de 500 vocábulos como o nível A+. A maior parte destes vocábulos pertencem ao vocabulário dos campos de número, bem como de trabalho e emprego, de comida e refeição, bem como de desporto, de animais, de corpo, de sentimento, de vida quotidiana, e de bens de uso doméstico (Tabela 6).

tipo	palavras
trabalho e emprego	bancário, cozinheiro, empregado, empregar, fábrica, fazenda, firma, jornalista, mineiro, projeto, salário, vendedor, agenda, cliente, diretor, liquidação, negócio, operário, papelaria, relatório, tradutor, voluntário, etc.
comida e refeição	talher, verdura, amargo, apetite, apimentar, cenoura, cozer, feijão, frito, grelhar, lanchar, laranja, milho, picante, presunto, torta, vinagre, etc.
desporto	capoeira, judô, time, atletismo, basquete, beisebol, bilião, bola, campeão, copa, cruzamento, maratona, torcida, vôlei, etc.
animais e vegetais	bambu, zebra, cabra, carneiro, cigarra, coelho, leão, palmeira, pássaro, pato, planta, etc.
corpo	barba, barriga, boca, nariz, ombro, rosto, bico, corno, dedo, estômago, garganta, orelha, peito, pescoço, unha, etc.
sentimento	assustar, cuidado, duvidar, intenção, recear, satisfazer, surpreender, vergonha, coragem, hesitar, incrível, preocupação, queixo, zangar, etc.
vida quotidiana	caseiro, gás, limpeza, carregar, cerimônia, cofre, descanso, desconto, desligar, diária, diário, encomendar, endereço, enviar, escovar, feriado, etc.
bens de uso doméstico	borracha, caderno, cartão, cheque, despertador, espelho, lápis, rádio, selo, toalha, cobertor, escova, garrafa, lâmpada, louça, sabonete, tesoura, etc.

Tabela 6: Vocabulários do nível A+

Vejamos seguidamente o vocabulário do nível B. Primeiro, adicionaram-se 172 palavras do nível avançado da Lista de L2. Os vocabulários destacáveis são de sentimento, de relação pessoal, de política (Tabela 7).

tipo	palavras
sentimento	chorar, preguiçoso, tímido, horrível, cômodo, nervoso, motivo, terror, chato, animar, pensamento, alegria, apetecer, emocionante, sentimento, etc.
relação pessoal	responder, cuidar, amizade, sobrinho, festejar, desculpa, membro, acompanhar, ligação, dedicar, celebrar
política	público, imigração, rei, cidadão, emigração, paz, político, governo, desenvolvimento, emigrante, desenvolver, crime, fronteira, emigrar

Tabela 7: Vocabulários do nível B que foram do intermédio e avançado

A seguir, à semelhança do que se pode verificar no nível A, podemos identificar vários vocábulos que não aparecem nos livros didácticos. De 108 palavras de diversos tipos, identificam-se os vocabulários relaccionados com sentimento, política e sociedade, nomes de locais sobretudo da Europa, e o português europeu.

tipo	palavras
sentimento	tranquilidade, engraçado, relaxar, positivo, paixão, sentimento, stress
política e sociedade	comunidade, urbano, rural, tradição, desenvolvimento, emigrante, desenvolver, crime, fronteira, emigrar
locais	Macau, Checa, Roma, barcelona, romano, Praga, Braga, Bruxelas
português europeu	bocado, fado, queima, fixe

Tabela 8: Vocabulários do nível B que não apareceram nos livros didácticos

Em suma, no decorrer da investigação, identifica-se mais de 1000 vocábulos do nível B que não aparecem na Lista L2 e se usam apenas num dos livros didácticos.

4.3. Problemas

A lista final deste artigo amplia e melhora a qualidade da Lista L2 de Torigoe (2016a). Porém, ainda existem diversos pontos que devem e podem ser melhorados. Por exemplo, há algumas palavras que parecem intuitivamente fundamentais mas ainda não estão classificadas nem no nível A, nem no B, conforme o critério adoptado neste artigo. Alguns dos exemplos são, *utilização*, *mundial*, *prisão*, *adversário*, *violência*, *relacionar*, *ocidental*, *claramente*, *sobreviver*, *imaginação*, *dependente*, etc.

Um outro ponto é que é, às vezes, arbitrária a diferença do grau de dificuldade entre alguns vocabulários do nível A+ e o do nível B, como seguidamente demonstrarei na Tabela 9.

nível A+	nível B
aposentar, artesanato, áspero, derrubar, enjoo, liquidação, retirar, tossir, etc.	diferença, caminhar, destino, tradicional, rir, suficiente, fresco, chorar, especial, paisagem, etc.

Tabela 9: Comparação de vocábulos entre o nível A+ e B

Assim como no procedimento da Lista L2, os vocabulários entre cada nível foram identificados com o critério mais objectivo possível, mas ainda é, de algum modo, arbitrário. Devido ao tamanho dos corpora de aprendentes e do corpus dos livros didácticos, a frequência do vocabulário torna-se menos de dez em cerca do primeiro 1000º vocábulo, o que não permite aplicar efectivamente o método estatístico como o exame de chi-quadrado às palavras de menos de 1000º lugar. Por este motivo, como o trabalho para o futuro, os corpora de aprendentes baseados no *QECR*, bem como o corpus de livros didácticos, devem ser ampliados, para analisar os dados de modo mais objectivo. A amplificação dos corpora também permite criar a lista de vocabulário do nível C.

Adicionalmente, neste artigo não se distinguem o português europeu e o do Brasil, devido à limitação nos dados disponíveis e à forma de ensino de português no Japão. Primeiro, até 2017, não é disponível nenhum corpus de aprendentes do português do Brasil. Segundo, as universidades do Japão oferecem o curso de "português" sem distinguir as variedades por motivo de diversas especializações dos professores e diversos interesses dos estudantes. Porém, é evidente a importância da diferença das duas variedades e é trabalho para o futuro a marcação de informação de variantes nas palavras específicas.

5. Conclusão e perspectiva para o futuro

Neste artigo, o autor criou e propôs a segunda versão-piloto do *Portuguese Vocabulary Profile*, a lista de vocabulário para aprendentes do PLE no Japão, a qual foi baseada nos copora de aprendentes e dos livros didácticos. A lista final deste artigo ampliou-se e também foi melhorada a qualidade do vocabulário do nível A (elementar) e B (intermédio), comparando com a lista do vocabulário de aprendentes que tinha sido criada em Torigoe (2016a). Porém, ainda persistem algumas problemas, os quais foram mencionados na secção anterior.

Igualmente, como um dos trabalhos futuros do projecto, o autor pretende adicionar a lista de colocações e expressões idiomáticas. Enquanto o *EVP* já integrou as expressões idiomáticas juntas às palavras, este trabalho, bem como Torigoe (2016a), elimina-as propositadamente para criar a lista de "palavras". Mas as expressões idiomáticas são essenciais para que os aprendentes prossigam para os níveis mais avançados (cf. Hunston & Francis 2000).

Para concluir, confirmarei o objectivo deste projecto: criar e propor um critério de vocabulário para aprendentes baseado nos dados objectivos, com o qual o autor deseja melhorar o ensino e materiais de português como língua estrangeira no Japão.

Agradecimentos

O presente estudo é financiado pela JSPS KAKENHI, Grant-in-Aid for Young Scientists (B) Grant Number: 15K16792. Igualmente, agradeço ao doutor professor Lúcio de Sousa (Universidade de Estudos Estrangeiros de Tóquio), ao professor Eliseu Pichitelli (Universidade de Estudos Estrangeiros de Tóquio) e à doutora professora Lúcila Etsuko Gibo (Universidade de Sofia, Tóquio) pela revisão do texto. Mas é o autor que assume toda a responsabilidade da informação no presente artigo.

Referências

Aires, Pedro. & Iyanaga, Shiro. (2010). *Verbos Fundamentais do Português*. CD-ROM Publicado privadamente.

Capel, Annette. (2010). A1-B2 Vocabulary: Insights and Issues Arising from the English Profile Wordlists Project. *English Vocabulary Journal*, Volume 1.

Capel, Annette. (2012). Completing the *English Vocabulary Profile*: C1 and C2 Vocabulary. *English Vocabulary Journal*, Volume 3, 1-14.

Council of Europa. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Extraído de: http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf.

Courses of Study. (2009). Tóquio: Ministério de Assuntos Internos e Comunicações do Japão. http://www.mext.go.jp/component/english/__icsFiles/afieldfile/2011/03/17/1303755_013.pdf (vista a 27 de Janeiro de 2017).

Davies, Mark. &. Preto-Bay, Ana Maria Raposo. (2008). *A Frequency Dictionary of Vocbulary*. New york: Routledge.

Hunston, Susan. & Francis, Gill. (2000). *Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English*. Amsterdam: John Benjamins Publishing.

Ishikawa, Shin-ichiro. (2006). *Vital Corpus 4500* (escrito em japonês). Tóquio: Bun-eido.

Leiria, Isabel. (2006). *Léxico, Aquisição e Ensino do Português Europeu Língua Não Materna*. Fundação Calouste Gulbenkian.

Ministério de Assuntos Internos e Comunicações do Japão. (2009). Population Census: Foreigners, by Nationality (11 Groups) Japan and Prefectures (escrito em japonês). Extraído de: http://www.e-stat.go.jp/SG1/estat/GL32020101.do?method=xlsDownload&fileId=000006911037&releaseCount=2.

Ministério de Ciência e Ensino do Japão. (2009). *Courses of Study: Secondary School Foreign Languages*. Extraído de: http://www.mext.go.jp/component/english/__icsFiles/afieldfile/2011/03/17/1303755_013.pdf.

Tono, Yukio. (2006). *Corpus Linguistics for Beginners* (escrito em japonês). Tóquio: Shogakukan..

Tono, Yukio. (2010). *Favourite Corpus 1800* (escrito em japonês). Tóquio: Tokyo Shoseki.

Tono, Yukio. (2013). *CEFR-J Handbook, A Resouce Book for Using CAN-DO Descriptors for English Language Teaching* (escrito em japonês). Tóquio: Taishukan.

Torigoe, Shintaro. (2015). Análise dos Níveis de Livros de Ensino da Língua Portuguesa através dos Descritores do CEFR-J. Apresentação Oral no 15º Encontro Nacional da Japan Association of Foreign Language Education.

Torigoe, Shintaro. (2016a). Seeking the Portuguese Vocabulary Profile: Pilot Study. In *EPiC Series in Language and Linguistics, CILC2016*.

Torigoe, Shintaro. (2016b). O Corpus e a Lista de Vocabulário de Livros Didácticos de Português. Apresentação Oral no Encontro Nacional da Associação Japonesa de Estudos Luso-Brasileiros.