A LEI PAZ-ANDRADE E O APROVEITAMENTO DO PORTUGUÊS PARA A TRADUÇÃO AUTOMÁTICA EM GALEGO

José Ramom Pichel, Iria de-Dios-Flores, Pablo Gamallo, Marco Neves*

Resumo

Para a tradução automática baseada em corpus, seja estatística ou neuronal, são necessários grandes volumes de traduções humanas entre duas línguas. Algumas línguas com poucos recursos não têm volumes suficientemente grandes, mas poderá haver acesso a recursos de línguas ou variedades vizinhas. Contudo, se existem entre ambas relações de parentesco controversas de uma perspetiva sociolinguística, reforçadas por quadros legais, poderão existir entraves institucionais ao desenvolvimento de tradutores automáticos oficialmente reconhecidos pelas autoridades linguísticas da língua com menos recursos. Neste artigo veremos como uma lei, que descongela as polémicas relações linguísticas galego-portuguesas (Lei Paz-Andrade), facilita legalmente o desenvolvimento de tradutores automáticos neuronais (NMT) com bons resultados para o galego, graças à utilização de corpora de português. Finalmente, propõe-se, a partir desta experiência, um método que pode ser aplicado a outras línguas com relações de parentesco controversas para o desenvolvimento de tradutores neuronais.

Palavras chave: Lei Paz-Andrade; legislação linguística; tradução automática; política linguística; galego; português.

THE PAZ-ANDRADE LAW AND THE USE OF PORTUGUESE FOR AUTOMATED TRANSLATION IN GALEGO

Abstract

In order to obtain high-quality corpus-based, statistical and neural machine translation, large volumes of human translations are needed between the two languages. Some low-resource languages do not have this volume of resources, but resources may be accessible from close-related languages or variants. Nevertheless, it is difficult to develop machine translation systems that are officially recognized by the language authorities of the language with fewer resources if there are sociolinguistic controversies with respect to their relatedness, further strengthened by legal frameworks. In this article, we will see how a law that untangles the controversial Galician-Portuguese linguistic relations (i.e., Paz-Andrade Law) legally facilitates the development of neural machine translation (NMT) systems with good results for Galician, thanks to the use of Portuguese corpora. Finally, based on this experience, we propose a method to develop neural translation systems that can be applied in the context of other low-resource languages that, despite being controversial, have a close relation with a language with greater resources.

Keywords: Paz-Andrade Law; linguistic legislation; automatic translation; language policy; Galician; Portuguese.

* José Ramom Pichel, investigador pós-doutorado em tecnologias da linguagem no Centro de Investigación en Tecnoloxías Intelixentes da Universidade de Santiago de Compostela (CITIUS), jramon.pichel@usc.es, 0000-0001-5172-6803

Iria de-Dios-Flores, investigadora pós-doutorado em tecnologias da linguagem no CITIUS da Universidade de Santiago de Compostela, iria.dedios@usc.es, 0000-0002-5941-1707

Pablo Gamallo, professor titular em tecnologias da linguagem no CITIUS da Universidade de Santiago de Compostela, pablo.gamallo@usc.gal, 0000-0002-5819-2469

Marco Neves, professor auxiliar na NOVA FCSH e investigador no CETAPS, mfneves@fcsh.unl.pt, 0000-0001-7648-9699

Artigo recebido em 10.06.2022. Avaliações às cegas: 26.08.2022 e 27.08.2022. Aceitação da versão final: 25.10.2022.

Citação recomendada: Pichel, José Ramom, de Dios Flores, Iria, Gamallo, Pablo, e Neves, Marco. (2022). A lei Paz-Andrade e o aproveitamento do português para a tradução automática em galego. *Revista de Llengua i Dret, Journal of Language and Law*, *78*, 35-55. https://doi.org/10.2436/rld.i78.2022.3845

Resumo

1 Introdução

2 Trabalhos relacionados

2.1 Língua, política e sociedade

2.1.1 A identidade linguística do galego: uma questão controversa

2.1.2 Políticas linguísticas na Galiza

2.2 Observações linguísticas

2.2.1 Padrão linguístico

2.2.2 Distância linguística

2.3 Tecnologias da linguagem: a tradução automática

2.3.1 SMT e NMT

3 Metodologia

4 Resultados

4.1 Tradutor estatístico de inglês-galego com recursos do português

4.2 Tradutor neuronal de espanhol-galego e de inglês-galego com recursos do português

5 Conclusões

6 Referências

1 Introdução

A Galiza e Portugal, que partilham uma língua desde a Idade Média (Maia, 1987), estão separados politicamente desde o século XII, embora sempre tenham mantido relações estreitas (Mattoso, 1986). Como resultado desta separação, foram iniciados processos institucionais, económicos, culturais e linguísticos diferentes entre a Galiza e Portugal. No que respeita aos aspetos linguísticos, tem havido desde o século XIX (Torres Feijó, 2019), e continua a haver, uma controvérsia entre os linguistas sobre se as duas variantes românicas faladas a norte (galego) e a sul (português) da fronteira são ou não a mesma língua. Este tema tornou-se premente com a mudança de regimes em Portugal e Espanha, em meados da década de 1970, e com as normas linguísticas previstas na Constituição espanhola e no Estatuto de autonomia galego. Estes dois quadros legislativos, com base na perceção social da separação linguística, ajudaram a congelar oficialmente as relações linguísticas galego-portuguesas até muito recentemente.

Assim, a Constituição espanhola de 1978, que estabeleceu um carácter obrigatório para o conhecimento do castelhano (Artigo 3.º), contribuiu para a substituição progressiva do galego como língua hegemónica pelo castelhano (Herrero Valeiro, 2003a), apesar de o galego ser na altura a principal língua ambiental na Galiza (Mira Pérez e Paredes, 2007). Já com a oficialização autonómica do galego em 1981, o galego acabou por se tornar, institucionalmente, uma língua equidistante entre o castelhano e o português, e, portanto, sem uma relação institucional especial com o português. Também contribuiu para este processo de distanciamento o facto de o Estatuto de Autonomia da Galiza não referir uma única vez Portugal ou a língua portuguesa, ao contrário do que aconteceu na Comunidade Autónoma da Extremadura e, em menor grau, na Comunidade Autónoma de Castela e Leão. No Estatuto de autonomia extremenho, Portugal é referido quatro vezes, a começar pelo primeiro artigo. No Estatuto de Autonomia de Castela e Leão, o artigo 66.º estabelece como prioridade a atenção a dar à relação com Portugal.

Por outro lado, duas leis europeias ajudaram a melhorar as relações galego-portuguesas: o espaço Schengen (Comissão Europeia, 2011) e a Carta Europeia das Línguas Regionais ou Minoritárias. No primeiro caso, a eliminação das estruturas fronteiriças entre Estados, como Espanha e Portugal, permitiu melhorar as relações económicas. No segundo caso, o apoio explícito da Europa à promoção das línguas regionais e minoritárias, como o galego, contribuiu para a defesa destes idiomas. Desta forma, como diz Piller (2016, p. 35) sobre a Carta Europeia das Línguas Regionais ou Minoritárias[1]: "Widely seen as one of the most progressive pieces of linguistic justice legislation globally, the Charter was designed to protect and promote regional and minority languages and to enable speakers to use them in private and public life". O desenvolvimento de tecnologias para estas línguas, entre outras questões, tornou-se uma necessidade, porque estas tornam mais fácil e mais barata a comunicação digital, pelo que tal desenvolvimento é recomendado pelos estados europeus que assinaram a Carta.

Após tentativas anteriores de estabelecer por lei relações privilegiadas entre a Galiza e Portugal (Pérez González e Martínez Puñal, 1986), em 2014, o Parlamento galego aprovou por unanimidade a Lei Paz-Andrade, que ajudou a potenciar as relações linguísticas entre galego e português, promovendo na Galiza uma relação especial com a língua portuguesa a nível institucional, educativo e audiovisual (em conformidade com a Diretiva 2007/65/CE). Anteriormente, em 2003, um novo padrão para o galego, mais convergente com o português, ajudou a esta aproximação. Este novo contexto e o novo quadro legislativo permitiram iniciar um novo processo de relações económicas, culturais e linguísticas entre a Galiza e Portugal, que também teve reflexos nas tecnologias linguísticas em geral e na tradução automática em particular.

Assim, neste artigo, veremos alguns exemplos de como a tradução automática em galego foi abordada durante o período de congelamento galego-português e como é abordada agora, após a Lei Paz-Andrade, com algumas experiências realizadas no âmbito do *Proxecto Nós*.[2] Finalmente, com base na experiência galego-portuguesa, propomos um método para projetos de desenvolvimento tecnológico, que pode ser aplicado noutros contextos europeus, tais como o romeno-moldavo, flamengo-holandês, servo-bósnio-croata, etc.

1 A Carta em versão inglesa.

2 www.nos.gal

2 Trabalhos relacionados

Nesta secção iremos rever os trabalhos e ideias particularmente relevantes para o objeto do nosso estudo. Está dividida em três blocos principais. Em primeiro lugar, na secção 2.1, faremos uma análise da situação do galego e da sua relação com a língua portuguesa de uma perspetiva sociolinguística, atendendo a casos de identidade e política linguística, bem como da Lei Paz-Andrade e de como ambas afetam o desenvolvimento de tecnologias da linguagem. Esta análise servirá para fornecer um quadro para a compreensão das reações aos projetos referidos na secção 2.3. Em segundo lugar, na secção 2.2, adotaremos uma perspetiva linguística para descrever as questões que ligam ou separam a língua portuguesa da língua galega, concentrando-nos, uma vez mais, no impacto destas questões no campo do processamento da linguagem natural. Isto irá determinar a viabilidade da utilização de recursos de uma língua para a outra. Finalmente, na secção 2.3, centrar-nos-emos no aspeto tecnológico do nosso objeto de estudo — a tradução automática — e nas diferentes tecnologias utilizadas na tradução automática em galego (RBMT, SMT e NMT).

2.1 Língua, política e sociedade

2.1.1 A identidade linguística do galego: uma questão controversa

Há em todo o mundo diferentes contextos em que existem dúvidas — e por vezes grandes controvérsias linguísticas — sobre se duas ou mais línguas são, de facto, línguas diferentes ou variantes diferentes da mesma língua (Rodrigues Fagim, 2001). Tal aconteceu historicamente, por exemplo na Europa, em línguas como o valenciano e o catalão (Flor, 2012), o flamengo e o holandês (Geerts et al., 1978), o moldavo e o romeno (Ciscel, 2006), as línguas escandinavas (Berg, 2016), ou, mais recentemente o servo-croata, dividido entre o bósnio, o croata e o sérvio, após a guerra dos Balcãs (Mønnesland, 1997).

Na Península Ibérica, temos o caso controverso da relação linguística entre o galego e o português. Assim, a consideração do galego como língua distinta do português, ou como uma das variantes da língua internacionalmente conhecida como português, é uma questão muito discutida (Carvalho Calero, 1990; Herrero, 2011; Ramallo e Rei-Doval, 2015). Os investigadores concordam num único ponto: na fase inicial da língua, que corresponde ao período de esplendor medieval tanto na Galiza como no Portugal atual, a língua era a mesma (Teyssier, 1982; Mariño Paz, 2008). Tradicionalmente, nos estudos linguísticos, nesta etapa da história, a língua românica é nomeada como *galego-português* (Maia, 1987), embora alguns autores recomendem o uso apenas do nome *galego* (Venâncio, 2020). A partir do século XV, alguns investigadores consideram que o galego e o português seguiram caminhos divergentes, resultando, na atualidade, em duas línguas próximas, mas diferentes (H. Monteagudo, 2017). Outros consideram que o galego e o português continuam a manter uma unidade estrutural suficiente para permitir que sejam considerados duas covariantes da mesma língua, como é o caso do português europeu e do português brasileiro, ou, no caso do espanhol, o espanhol de Espanha e o espanhol do México ou da Argentina (Henríquez Salido, 2021).

Estas duas conceções da língua na Galiza tomaram a forma de duas propostas diferentes para o padrão galego a ser estudado nas escolas galegas, uma vez formalizado um estatuto de autonomia para a Galiza em 1981 (Duarte Collazo, 2014). A perspetiva finalmente adotada foi a de que o galego e o português são duas línguas diferentes (Sánchez Vidal, 2005). Nesta conceção, o galego é uma terceira língua entre o português e o castelhano. Esta ideologia pode ser chamada de *autonomista* (Labraña e Vázquez, 1997), uma vez que concebe o galego como uma língua diferente da portuguesa, presente em diferentes regiões espanholas (Galiza, Astúrias, Leão e mesmo Extremadura). A outra ideologia, conhecida como *reintegracionista*, concebe o galego como uma das variantes da língua portuguesa (Pérez-Barreiro Nolla, 1990), devendo, portanto, convergir com o resto das variantes do português (por exemplo, Portugal, Brasil, PALOP).

A visão linguística autonomista, sendo a corrente que acabou por se tornar dominante, concebeu, do início da década de 1980 até hoje, políticas e estratégias que têm como prioridade a geração de recursos linguísticos próprios para uma língua com um padrão próprio entre o espanhol e o português. Esta conceção repercutiu-se no desenvolvimento de tecnologias linguísticas (Gómez Guinovart, 2006). Sendo possível a utilização livre dos recursos linguísticos do português com vista ao seu aproveitamento no galego, estes não eram oficialmente aceites por não terem o carimbo de autoridade dos organismos linguísticos galegos (Samartim, 2012). Tal

afetou os desenvolvimentos das tecnologias linguísticas em geral, e no presente caso, a tradução automática, pois não eram considerados como recursos confiáveis, visto não terem esse selo *oficial*.

2.1.2 Políticas linguísticas na Galiza

Existem diferentes leis linguísticas que se aplicam na Galiza, investigadas por diferentes autores e autoras (X. Monteagudo, 1990; Ochoa Monzó, 2006; Nogueira López, 2020; Pacho Blanco, 2016), que criam direta ou indiretamente um quadro de identidade cultural e linguístico não só de reconhecimento do galego, mas também de autorreconhecimento (Busto Miramontes, 2016), produzindo quadros de relações com outras comunidades autónomas ou com outros países (Carvalhido, 2017). Estas leis podem ser agrupadas nos seguintes períodos em relação ao galego: período de proteção do galego e período de proteção do galego e línguas estrangeiras. No que diz respeito à relação entre o galego e o português, como se segue: período de congelamento e período de descongelamento galego-português. Estes períodos estão representados na Tabela 1.

Tabela 1. Períodos linguísticos na Galiza e em relação a Portugal

	1978–2009	**2009–2014**	**2015–2022**
Proteção do galego			
Proteção do galego e línguas estrangeiras			
Congelamento galego-português			
Descongelamento galego-português			

No período de proteção do galego, existem diferentes leis, entre as quais destacamos, em primeiro lugar, o Artigo 3.º da Constituição espanhola de 1978 (Española, 1978), que, tal como o Artigo 4.º da Constituição espanhola de 1931 (Corcuera Atienza, 2000), oficializam o castelhano como a única língua do Estado, obrigando os cidadãos galegos a conhecê-la:

> Artículo 3. El castellano y las demás lenguas españolas
>
> 1. El castellano es la lengua española oficial del Estado.
> 2. Todos los españoles tienen el deber de conocerla y el derecho a usarla. Las demás lenguas españolas serán también oficiales en las respectivas Comunidades Autónomas de acuerdo con sus Estatutos.
> 3. La riqueza de las distintas modalidades lingüísticas de España es un patrimonio cultural que será objeto de especial respeto y protección.
>
> Artículo 4. El castellano y las demás lenguas españolas
>
> 1. El castellano es el idioma oficial de la República. Las demás lenguas españolas serán también oficiales.
> 2. Todo español tiene la obligación de saberlo y derecho de usarlo, sin perjuicio de los derechos que las leyes del Estado reconozcan a las lenguas de las provincias o regiones.
> 3. Salvo lo que disponga en leyes especiales a nadie se le podrá exigir el conocimiento ni el uso de ninguna lengua regional.

Por outro lado, tendo em conta que a maioria da cidadania galega adulta na década de 1980 se expressava em galego (González González, 1985), existem duas leis autonómicas que definem a política linguística da comunidade: o Estatuto de Autonomia (1981) e a Lei de Normalização Linguística (1983). No Estatuto de Autonomia, afirma-se o direito de conhecer o galego no:

> Artigo 5
>
> 1. A lingua propia de Galicia é o galego.

2. Os idiomas galego e castelán son oficiais en Galicia e todos teñen o dereito de os coñecer e de os usar.

3. Os poderes públicos de Galicia garantirán o uso normal e oficial dos dous idiomas e potenciarán o emprego do galego en tódolos planos da vida pública, cultural e informativa, e disporán os medios necesarios para facilita-lo seu coñecemento.

4. Ninguén poderá ser discriminado por causa da lingua.

Com o Artigo 4.º da Lei da Normalização Linguística (1983), a administração local na Galiza passa a ter também, a partir dessa data, o galego como língua oficial e os topónimos recuperam as suas formas originais galegas. As questões linguísticas relacionadas com a administração local seriam desenvolvidas posteriormente na Lei da Administração Local de 1997 (Lei 5/1997):

Artigo 4

1. El gallego, como lengua propia de Galicia, es lengua oficial de las instituciones de la Comunidad Autónoma, de su Administración, de la Administración Local y de las Entidades Públicas dependientes de la Comunidad Autónoma.

2. También lo es el castellano como lengua oficial del Estado.

Para além destas leis obrigatórias, durante o período de proteção do galego, aparece a Carta Europeia das Línguas Regionais ou Minoritárias como convenção internacional, ratificada pelo Reino de Espanha, estando, portanto, também em vigor na Comunidade Autónoma da Galiza a partir do dia 1 de agosto de 2001. A Carta faz várias recomendações, com um seguimento, nessa altura bianual, pelo Secretário-Geral do Conselho da Europa, entre as quais se salienta a promoção da utilização das línguas minoritárias na vida pública, especialmente na educação, justiça, administração pública e meios de comunicação social, entre outros sectores[3]. Por fim, o período de promoção do galego termina com o Decreto 124/2007, do governo bipartido PSOE-BNG (Lastra Muruais, 2010), que abria legalmente a possibilidade de que 50% das disciplinas de ensino obrigatório fossem lecionadas em galego. Posteriormente, no ano 2010, surgiu o Decreto 79/2010 com o novo governo do PP, que, em nome da promoção do multilinguismo, inicia o período, ainda em vigor, da proteção das línguas estrangeiras no ensino, nomeadamente o inglês. A principal novidade deste decreto foi a possibilidade de consultar as famílias que têm filhos na mesma turma sobre a língua em que querem que os seus filhos sejam educados. O inquérito às famílias tem uma série de semelhanças com o que mais tarde ficou conhecido como o *pin parental* (Climent Gallart, 2020, p. 4). Segundo descreve o autor no artigo, e que traduzimos para português, o *pin parental* é:

> Um pedido dirigido aos diretores das escolas. Este pedido pede à direção da escola que informe previamente os pais ou tutores, através de uma autorização expressa, sobre qualquer assunto, conversa, seminário ou atividade que afete questões morais ou sexualmente controversas socialmente, que possam ser intrusivas para a consciência e privacidade dos seus filhos, para que, como pais, possam conhecê-los e analisá-los previamente, refletir sobre eles e, nessa base, dar o seu consentimento — ou não — para que os seus filhos frequentem essa formação.

Tal facilitou, de alguma forma, a equiparação no ensino obrigatório das línguas estrangeiras não nomeadas nem na Constituição, à língua oficial do Reino de Espanha e à língua própria da Comunidade Autónoma da Galiza (Villares Naveira, 2010). Como resultado deste período de proteção de línguas estrangeiras, de acordo com Ramallo (2016, p. 8), o quarto relatório do Comité de seguimento da Carta Europeia das Línguas Regionais ou Minoritárias do Conselho da Europa de 2016, em relação a este decreto, demonstra preocupação ao descrever:

> Tamén hai motivos evidentes de alarma na situación das linguas oficiais no sistema educativo, en concreto o galego en Galicia e o catalán en País Valenciano e Illas Baleares. Nestes tres territorios, hai en marcha programas trilingües no sistema educativo, o que supón unha menor presenza das linguas minoritarias.

3 https://www.coe.int/en/web/compass/european-charter-for-regional-or-minority-languages

Além dos períodos identificados para o galego, existem dois períodos relativos às relações controversas entre o galego e o português: aquele que podemos chamar de período de congelamento galego-português e o período de descongelamento galego-português, que teve início com a Lei Paz-Andrade e que detalharemos em seguida.

2.1.2.1 Lei Paz-Andrade para o aproveitamento da língua portuguesa

Em 2014, foi aprovada a chamada Lei Paz-Andrade (Morell, 2019), que encetou na Galiza o reencontro com o português, suspenso desde o início da década de 1980. Esta lei, inicialmente proposta por uma Iniciativa Legislativa popular com mais de 17 000 assinaturas (Miranda Gonçalves, 2016), foi aprovada por unanimidade pelos grupos políticos. A lei resultou do facto de a ideia de uma relação mais estreita com Portugal através da língua nunca se ter desvanecido na Galiza (Fernández Souto et al. 2019; Quiroga, 2013).

Esta lei centra-se em diferentes áreas que permitem melhorar as relações culturais e económicas entre a Galiza e Portugal, através do conhecimento do português. Tal reflete-se no início dos artigos da Lei Paz-Andrade:

> A lingua propia de Galicia, polo feito de ser intercomprensible co portugués, outorga unha valiosa vantaxe competitiva á cidadanía galega en moitas vertentes, nomeadamente na cultural pero tamén na económica. Por isto debemos dotarnos de métodos formativos e comunicativos que nos permitan desenvolvernos con naturalidade nunha lingua que nos é moi próxima e nos concede unha grande proxección internacional.

As principais prioridades da Lei Paz-Andrade são as seguintes:

- Promoção do ensino do português na educação obrigatória e na sociedade galega em geral, com especial ênfase na função pública.
- Relações estratégicas entre instituições galegas e portuguesas.
- Promoção de fóruns e eventos para fomentar as relações económicas, culturais, ambientais, desportivas, etc. entre a Galiza e Portugal.
- Fomentar o intercâmbio e colaboração em produções audiovisuais galego-portuguesas, com especial atenção ao intercâmbio e colaboração em conformidade com a Diretiva 2007/65/CE do Parlamento Europeu e do Conselho, relativa aos serviços de comunicação social audiovisual sem fronteiras.

Finalmente, a Lei Paz-Andrade (Lei 1/2014) significou uma mudança não só para a Galiza, mas também para Portugal (Rodríguez Fernández, 2018), uma vez que uma comunidade autónoma espanhola, com a categoria de nacionalidade histórica, se mostrou interessada em dar prioridade à relação linguística, cultural e económica com Portugal. Esta lei, para além das suas repercussões legislativas e económicas, promoveu uma melhor relação entre o galego e o português, o que, como veremos abaixo, na área das tecnologias linguísticas, acelerou a aquisição de recursos do português para serem reutilizados pelo galego, especialmente no campo da tradução automática.

2.2 Observações linguísticas

Após analisarmos as questões relacionadas com a sociolinguística e as políticas linguísticas, comentaremos a relação entre o galego e o português a partir de dois aspetos linguísticos: a evolução do padrão galego em relação ao português e a distância entre os padrões galego e português. No primeiro caso, veremos como as escolhas linguísticas de um padrão de galego ou outro influenciam a aproximação ou o distanciamento em relação ao português. No segundo caso, veremos como existem métricas para o cálculo automático da distância, o que é importante para saber se podemos utilizar na tradução automática baseada em corpus (estatística e neural) recursos linguísticos da língua com maiores recursos para uso com a língua com menos recursos, como no caso do português e do galego.

2.2.1 Padrão linguístico

O padrão escrito do galego teve dois grandes períodos desde a década de 1980, quando o galego foi incorporado no ensino obrigatório: menor convergência (1981–2003) e maior convergência (2003–presente) com o padrão português.

Desde a entrada em vigor do Estatuto de Autonomia, o galego tem sido progressivamente incorporado na administração galega e na educação obrigatória através das leis acima descritas. A introdução do galego na educação tornou necessário criar um padrão, que, devido à grande proximidade linguística galego-portuguesa, poderia convergir mais ou menos com o português (Alonso Pintos, 2005). Como comentamos, estas possibilidades geraram diferentes controvérsias, evoluções e ruturas no padrão galego (Salgado e Monteagudo, 1993; H. Monteagudo, 1990), que resumiremos em seguida, mas que já existiam anteriormente (Montero Santalha, 2001).

Em 1980, uma comissão dirigida pelo Professor Ricardo Carvalho Calero propôs uma norma que mantinha a unidade linguística galego-portuguesa, com uma ortografia próxima à castelhana para o galego, mas com algumas características comuns ao português (por exemplo, pronomes enclíticos ligados por hífenes ao verbo) (Mariño Paz, 2002). Esta norma não foi adotada para a educação pública porque, entre o final de 1982 e o início de 1983, foi acordado um novo padrão denominado *Normas ortográficas e morfolóxicas do idioma galego*, as quais, como indica Regueiro Tenreiro, foram "fruto de un acuerdo del Instituto de la Lengua Gallega y de la Real Academia Galega que por ley, Ley de Normalización Lingüística, es la depositaria de la responsabilidad de cuanto se refiere a actualización y normativa lingüística" (1994, p. 41). Por um lado, estas novas regras definiram a identidade do galego como uma nova língua românica separada do português e do castelhano e, por outro lado, iniciaram o caminho para a construção de uma língua galega com padrão autónomo (de forma significativa com a mesma grafia do castelhano, exceto alguns traços como a simplificação de *g*, *j* e *x* num único símbolo ortográfico *x*, etc.) e promovendo a morfologia comum galego-portuguesa nalguns traços (por exemplo, plurais em *-ais*, morfologia *-aría* e *-abel*).

A partir desse momento, houve diferentes evoluções (Muñoz Carrobles, 2008), que poderíamos sintetizar em duas grandes visões (Conde, 2003), embora com muitas subvariantes (Álvarez Cáccamo e Herrero Valeiro 1996), que não aceitaram esta identidade linguística separada do português e esta ortografia posterior. A primeira foi reconhecida como *reintegracionista de mínimos*, cujos defensores assumiram como norma um galego com traços mais convergentes com o português, mas com ortografia na sua maioria convergente com o castelhano. A segunda ficou conhecida como *reintegracionista de máximos* (Duarte Collazo, 2014), e foi liderada por Carvalho Calero numa nova associação chamada AGAL (Associaçom Galega da Língua), que defende a identidade linguística galego-portuguesa e a utilização de uma norma muito próxima ou idêntica à norma portuguesa europeia para o galego escrito.

Durante este período, as crianças galegas aprenderam na escola que, para além das gralhas comuns no galego — vulgarismos, hiperessencialismos ou castelhanismos, nomenclatura que também existe nas outras línguas minoritárias do Reino de Espanha —, havia uma nova categoria que também interferia no equilíbrio do padrão do galego separado: os *lusismos* (Silva Domínguez, 2008). Um *lusismo* é, de acordo com a Real Academia Galega "un vocábulo ou modo de falar propio do portugués empregado noutra lingua"[4]. No nosso estudo de caso, em galego.

Este período do galego separado menos convergente com o português durou entre 1982 e 2003, quando as posições *oficial* e de *mínimos reintegracionistas* chegaram a um acordo no qual, sem modificar a identidade do galego como língua equidistante entre castelhano e português, foram adotadas opções mais convergentes com o português de Portugal. A partir deste acordo, que deu lugar a uma nova norma do galego (Instituto da Lingua Galega, 2004), o galego, mantendo a sua identidade linguística separada do português, aproximou-se mais do seu padrão em certos aspetos lexicais e morfológicos. Assim, deu-se prioridade às formas lexicais e morfológicas mais convergentes com as portuguesas, eliminando em alguns casos formas mais distantes. No campo lexical, podemos destacar: *rector* por *reitor*, *secta* por *seita* ou *estudiar* por *estudar*. No campo morfológico, os sufixos *-bel/aría* são equiparados a *-ble/ería*, dando assim como válido: *estable* e

4 https://academia.gal/dicionario/-/termo/busca/lusismo

estábel, *librería* e *libraría* (dando prioridade a esta última) ou *admirable* e *admirábel*. Também nos grupos consonânticos, a eliminação dos grupos *-ct* e *-cc-* após *i* ou *u* levou à substituição de *producto* por *produto* ou *construcción* por *construción*. Finalmente, vale a pena notar que este acordo se tornou muito visível no discurso quotidiano por causa de duas palavras de uso frequente: *grazas* e *servizo*, que substituíram *gracias* e *servicio*. Também deve ser dada especial atenção ao topónimo da comunidade autónoma, que também está incluído no presente acordo. Assim, a partir deste acordo, a forma *Galiza*, a forma mais frequente desde o nascimento da língua na Idade Média e ainda a única utilizada em Portugal (Montero Santalha, 2008), foi também reconhecida como uma forma válida dentro do galego. No entanto, o nome oficial permaneceria inalterado e continuaria a ser *Galicia* em galego.

A partir da norma do ano 2003, apenas as organizações reintegracionistas, que veem o galego e o português como duas variantes da mesma língua, são deixadas de fora. Contudo, parece que já não aspiram, exceto a Associaçom de Estudos Galegos (Henríquez Salido, 2020; Basanta e Regueira, 2017), a que a variedade galega da língua tenha uma subnorma própria no português, em coordenação com as outras subnormas, principalmente português europeu e português do Brasil (Henríquez Salido, 2021).

Para concluir esta secção, é importante salientar que, no galego autonómico, pode haver formas lexicais, morfológicas e sintáticas mais próximas do castelhano ou do português. Por exemplo, o jornal digital galego *Nós*[5] usa as denominações das feiras para os dias da semana, como em Portugal, enquanto o jornal digital Praza[6] usa *luns*, *martes*, *mércores*, etc. — sendo ambas as opções válidas. Dependendo da instituição, empresa ou associação, algumas formas podem ser mais aceites do que outras e tal representa um desafio para a concretização, por exemplo, da tradução automática neuronal. O corpus de treino condiciona o resultado que o tradutor automático neuronal aprende. É por esta razão que utilizar corpora de português no quadro das polémicas relações galego-portuguesas pode criar dificuldades na aceitação oficial de sistemas de tradução automática neuronal (NMT).

2.2.2 Distância linguística

Nesta secção introduzimos o conceito de distância linguística, que será de particular relevância para a metodologia apresentada abaixo. De acordo com Pichel Campos (2020, p. 33):

> As línguas têm sofrido alterações ao longo da sua história, tanto interna como externamente, em relação a outras línguas. A fim de medir esta evolução, foram propostas abordagens diferentes a partir de estudos filogenéticos, na dialetologia ou na área da aquisição de segunda língua. No domínio do processamento de línguas naturais, este papel tem cabido à identificação automática das línguas e à distância entre línguas.

Podemos, portanto, dizer, de forma resumida, que existem métodos provenientes de estudos linguísticos, como a filogenética (Petroni e Serva, 2010), a dialetologia computacional (Nerbonne e Heeringa, 1997), a aprendizagem de línguas (Chiswick e Miller, 2005), o estudo diacrónico sobre a evolução das línguas (Lai et al., 2018) ou a dialetologia (Lui e Cook, 2013), e outros que partem da matemática como a identificação automática de línguas e variantes (Jauhiainen et al., 2019; Zampieri et al., 2015; Molina et al., 2016), os estudos de economia (Isphording e Otten, 2013), os estudos sobre a dinâmica da sobrevivência linguística (Mira e Paredes, 2005) o estudo sobre a inteligibilidade mútua entre línguas (Moberg et al., 2007) ou a distância entre línguas (Pichel Campos, 2020).

Com vista a criar um tradutor para uma língua com poucos recursos, é interessante ver a que distância está esta língua de outra língua com mais recursos. Se a distância entre as duas for próxima ou muito próxima, talvez seja possível utilizar os recursos da língua que tem mais recursos na que tem menos recursos. Para o fazer, teremos de realizar diferentes transformações que dependerão de cada par de línguas. Na Secção 4 de Resultados, explicaremos algumas transformações possíveis no caso do português e do galego, que é o tema principal do nosso artigo.

5 https://www.nosdiario.gal/

6 https://praza.gal/

Assim, no nosso caso particular, uma vez que os sistemas de tradução automática estatística e neuronal utilizam principalmente grandes corpora de traduções humanas, utilizaremos uma métrica de distância linguística entre galego e português baseada na métrica da *perplexity* (Gamallo et al., 2017; Pichel Campos, 2020), porque calcula especificamente a distância entre corpora de duas línguas.

Como a distância em *perplexity* é muito próxima entre português e galego[7] (distância de 5,84 com o galego, em comparação com 7,72 com o castelhano), os corpora do português podem ser transformados em corpora de galego através de diferentes transformações lexicais e ortográficas, permitindo obter tradutores automáticos para o galego, que, de outra forma, seriam muito dispendiosos em termos de tempo e orçamento.

Por exemplo, podemos transformar aspetos linguísticos e até ortográficos da língua com mais recursos (português) para a língua com menos recursos (galego), mudando a grafia portuguesa *nh* e *lh* e *ñ* e *ll* do galego, por exemplo, de *caminho* para *camiño* e de *coelho* para *coello*, ou mudando o léxico português *carro* para *coche* em galego.

2.3 Tecnologias da linguagem: a tradução automática

Embora existam diferentes aplicações das tecnologias linguísticas para o galego, centrar-nos-emos no caso particular da tradução automática, uma vez que a utilização de recursos linguísticos do português para o galego pode ser especialmente relevante. Há duas abordagens principais à tradução automática: uma baseada em dicionários e regras gramaticais e outra baseada em traduções humanas e aprendizagem automática a partir destes dados.

A primeira é conhecida como tradução automática baseada em regras, *Rule-based Machine Translation* (RBMT) (Koehn, 2009); a segunda era inicialmente conhecida como tradução automática baseada em estatística (SMT), utilizando agora técnicas de aprendizagem profunda, como tradução automática neuronal, isto é, *Neural Machine Translation* (NMT) (Bahdanau et al. 2015). Atualmente, o estado da arte em tradução automática é a tradução automática neuronal (NMT). No entanto, existem contextos nos quais a tradução automática com sistemas RBMT continua a ser a que obtém os melhores resultados, como é caso da tradução de galego, se estiver a ser realizada com variantes românicas próximas, como o português ou o castelhano (Bayón e Sánchez-Gijón, 2019). O mesmo parece não se aplicar, de acordo com os resultados iniciais, à tradução automática entre o galego e outras línguas mais distantes.

Em seguida, iremos rever todos estes sistemas para o galego. Para a RBMT a partir do galego ou para o galego, foram criadas diferentes plataformas desde finais da década de 1990 (Gómez Guinovart, 1997), quer em software proprietário quer em software livre, especialmente para línguas próximas (português e castelhano) e para o inglês. Assim, foi desenvolvido um tradutor RBMT de alta qualidade chamado *ES-GL* (González González, 2006), que não só trabalhou aspetos léxicos, sintáticos, mas também semânticos. Este sistema foi desenvolvido em conjunto pela empresa Lucy Ibérica, uma subsidiária na Península Ibérica da Lucy Software, e o Centro galego de Investigação em Humanidades Ramón Piñeiro. Posteriormente, o projeto *Opentrad* (Alegría Loinaz et al., 2006), desenvolvido por várias universidades (Universitat d'Alacant, Universitat Politècnica de Catalunya, Euskal Herriko Unibersitatea - Universidad del País Vasco e Universidade de Vigo) e várias empresas (imaxin|software, Elhuyar Fundazioa e Eleka (Rodríguez González et al., 2016)), gerou dois motores de tradução automática de código aberto chamados *Apertium*[8] (Forcada et al., 2011), inicialmente concebido para línguas próximas, e *Matxin*[9], concebido para um par de línguas distantes como o espanhol e o basco.

Posteriormente, foram desenvolvidas várias melhorias no *Apertium*, o que permitiu a sua evolução para uso com línguas distantes. Também uma comunidade internacional de código aberto desenvolveu novos pares de línguas, com mais de 30 pares de línguas estáveis de todo o mundo. Finalmente, podemos destacar, para o galego, que o sistema de tradução da Xunta de Galicia, chamado *Gaio*[10], e o sistema de tradução da Secretaria-

7 https://gramatica.usc.es/~gamallo/php/distance/index.php

8 https://github.com/apertium

9 http://matxin.elhuyar.eus/

10 http://tradutorgaio.xunta.gal/TradutorPublico/traducir/index

Geral da Administração Digital de Espanha (*PLATA*), se baseiam em personalizações e melhorias deste sistema desenvolvido pela empresa galega imaxin|software (Pichel Campos et al., 2014).

2.3.1 SMT e NMT

Por fim, falaremos sobre os sistemas de tradução SMT e NMT, que se baseiam sobretudo em corpora de traduções humanas. Os primeiros utilizam sistemas estatísticos para gerar traduções e os segundos utilizam sistemas de aprendizagem profunda (*deep learning*), tais como redes neuronais artificiais (Koehn, 2020).

Entre os protótipos, podemos destacar o sistema chamado *Carvalho*, desenvolvido em 2008 pela imaxin|software, que publicou os seus resultados em 2009 e 2010 (Pichel Campos et al., 2010). Para a tradução entre inglês e galego, a empresa utilizou corpora portugueses e transformou-os em galego. Para tal, traduziu os corpora de português para o galego utilizando uma versão especial de *Apertium* (do português para o galego) desenvolvida pela imaxin|software e transliterando posteriormente as palavras desconhecidas com um transliterador chamado *por2gal*[11]. Este script converte a ortografia do português europeu para a ortografia atual do galego, muito próxima da castelhana.

O *Google Translate* incluiu, no ano 2009, o galego no seu sistema de tradução automática[12]. Nesta primeira versão, embora não nas atuais, há indícios de que tenha utilizado a mesma estratégia do *Carvalho* (Campos, 2019), para além de corpora de castelhano e português do Brasil. No entanto, nem sempre foi utilizada a transliteração, uma vez que manteve o uso da ortografia portuguesa, por exemplo na não eliminação do hífen para os pronomes enclíticos. Outros sistemas também incorporaram mais tarde o galego, tais como o da Microsoft[13] e o da Yandex[14]. Posteriormente, foram desenvolvidos diferentes protótipos em sistemas de aprendizagem profunda (NMT) que também utilizaram esta estratégia para o galego. Diferentes protótipos estão a ser desenvolvidos no âmbito do *Proxecto Nós*[15] pelo CITIUS que também utilizam, neste caso para o par espanhol-galego, as mesmas técnicas realizadas em SMT, cujos principais resultados veremos mais tarde.

3 Metodologia

Seguidamente, propomos uma metodologia para que uma língua com poucos recursos possa obter e transformar corpora a partir de uma língua muito próxima que disponha de recursos suficientes. É importante notar que entre as duas existiram ou continuam a existir relações controversas de parentesco. Estes corpora transformados serão utilizados para o treino de sistemas de tradução automática baseados em corpora (SMT e NMT). Esta metodologia tem sido aplicada ao galego e ao português para o desenvolvimento de tradução automática (SMT e NMT) para o galego com base em corpus a partir de recursos do português. Neste caso, o galego tem muitos menos recursos do que o português, há uma distância linguística próxima medida em *perplexity* e têm uma relação controversa de parentesco.

É de salientar que este método foi concebido especialmente para sistemas de tradução automática NMT (embora também funcione para SMT, que já não é o estado da arte), pois estes precisam de enormes quantidades de corpora de tradução humana entre línguas (corpora paralelos). Isto não invalida outras estratégias para a construção destes tradutores, tais como não a utilização de corpora paralelos, mas comparáveis (Artetxe et al., 2017), ou a utilização de corpora polissintéticos (Ortega et al., 2020).

A metodologia é descrita e explicada a seguir:

1. **Medir distância linguística:** calcular a distância linguística baseada em corpus (distância PLD baseada em *perplexity*) entre a língua ou variedade com mais recursos e a língua ou variedade com menos recursos. Se a distância for próxima ou muito próxima, a utilização de corpora da língua

11 https://github.com/gamallo/port2gal

12 https://ctnl.gal/web/notic.php?ide=301&desc=o_google_translate_incorpora_o_galego

13 https://www.bing.com/translator?from=gl&to=mg&setlang=sw

14 https://translate.yandex.com/?lang=gl-th&text=

15 https://www.lingua.gal/recursos/todos/_/promovelo/contido_607/nos-intelixencia-artificial-servizo-lingua-galega

com mais recursos pode ser adequada para a língua com menos recursos. Caso contrário, verificar se existe alguma relação de parentesco que relacione as línguas ou variedades (não é necessário haver métricas matemáticas de cálculo entre línguas, mas sim haver estudos de linguistas que atestem a existência de um parentesco, como acontece com o bretão e o galês).

2. **Verificar controvérsia entre as línguas:** investigar se há controvérsia quanto à relação de parentesco entre línguas. Caso haja controvérsia (por exemplo, galego/português, valenciano/catalão, romeno/moldavo, servo-croata ou sérvio/croata/bósnio), observar o atual quadro legislativo em relação ao congelamento ou descongelamento legislativo do parentesco entre a língua com menos recursos e a língua com mais recursos (por exemplo, o galego e o português), pois tal indica a existência de uma abertura política que atenuaria uma possível rejeição de formas próprias da variedade com mais recursos.

3. **Decidir sobre o tipo de desenvolvimento do tradutor automático:** após verificar a distância linguística e o estado atual dos quadros legislativos, decidir sobre o tipo de desenvolvimento a ser realizado em tradutores automáticos baseados em corpus:

    a. Negativo: a distância linguística entre as línguas é elevada ou não há indicação de uma relação linguística. Neste caso, nenhum desenvolvimento é desejável.

    b. Protótipo: se houver distância linguística pouco elevada (*perplexity*) ou indícios de relação linguística, e o quadro legislativo não for favorável, desenvolver um protótipo não oficial, tendo em conta que é muito provável que os organismos linguísticos não permitam que o tradutor automático tenha estatuto oficial, uma vez que se baseia numa língua relacionada com a qual existe controvérsia. Ir para o ponto 4 da metodologia.

    c. Demonstrador: se a distância linguística estiver próxima e o quadro legislativo for favorável, desenvolver um demonstrador. Deve notar-se que os resultados podem ser inicialmente relevantes, daí o termo *demonstrador*. Ir para o ponto 4 da metodologia.

4. Começar a recolher corpora linguísticos da língua com mais recursos para serem transformados em corpora na língua com menos recursos, a fim de treinar os sistemas de tradução automática SMT/NMT.

5. Finalizar a metodologia.

4 Resultados

A seguir explicaremos os diferentes resultados da aplicação desta metodologia no desenvolvimento de sistemas de tradução automática estatística (SMT) e neuronal (NMT) para o galego reutilizando recursos linguísticos do português europeu.

O primeiro resultado está relacionado com a fase de Protótipo da metodologia, uma vez que, na época a que se reporta (2009), existia um quadro legal de congelamento de relações entre o galego e o português. Também não existiam medições de distância linguística baseadas em corpus, embora os linguistas defendessem uma relação especial entre as duas variantes. O dispositivo de tradução gerado foi um protótipo SMT inglês-galego, que era o estado da arte. Em contraste, o segundo resultado, a que podemos chamar Demonstrador, parte de uma distância linguística já calculada, que mostrava grande proximidade entre o português e o galego; a distância foi calculada utilizando uma métrica baseada em *perplexity*; existia também um quadro legislativo favorável.

Em relação à distância linguística, foi realizado o cálculo da PLD, que vai de 0 a N, sendo a PLD < 8 o valor de duas línguas/variedades próximas. Quanto maior for o valor de PLD, maior a distância entre as línguas comparadas (Pichel Campos, 2021). A distância exata entre o galego e o português, utilizando uma ortografia artificial comum construída ad hoc, é PLD: 5,47 (Pichel Campos, 2020). Por uma questão de contraste, é importante notar que a distância entre duas línguas próximas como o português e o espanhol é PLD: 7,49.

No que diz respeito a quadros legislativos, é importante notar que a Lei Paz-Andrade descongelou as relações linguísticas galego-portuguesas, o que permitiu o desenvolvimento *oficial* de recursos para o galego através da transformação dos recursos do português. Tal torna muito mais fácil a procura de recursos para a tradução automática neuronal para o galego. Por fim, apresentaremos os resultados de um tradutor de máquina neuronal de espanhol-galego e inglês-galego construído a partir da transformação de corpora de espanhol-português e inglês-português no âmbito do *Proxecto Nós*. Estes resultados são parte de um processo de criação de um futuro Demonstrador.

4.1 Tradutor estatístico de inglês-galego com recursos do português

Em relação ao tradutor SMT, foi desenvolvido o *Carvalho SMT* (Campos, 2009; Fernández Malvar et al., 2010), um protótipo de um tradutor estatístico de inglês-galego que utilizava Moses (Koehn et al., 2007) e o corpus de inglês-português *Europarl* para o treino (Koehn, 2005).

Seguindo a metodologia, nessa altura não havia nenhuma métrica quantitativa extraída de corpora (Gamallo et al. 2017) com base em *perplexity*, nem havia nenhum quadro legislativo favorável à utilização dos recursos do português para o galego. No entanto, teorias linguísticas desenvolvidas por diferentes autores (por exemplo, Carvalho Calero, 1985; Torres Feijó, 2000; Rodrigues Fagim, 2001; Herrero Valeiro, 2003b; Montero Santalha, 2010) levaram à conclusão de que poderia ser interessante experimentar esta abordagem inovadora no campo das tecnologias linguísticas. Por conseguinte, a nossa metodologia aconselha a geração de um Protótipo. Para este fim, os corpora de inglês-galego utilizados para treinar o sistema de tradução estatística Moses foram obtidos de corpora de inglês-português transformados utilizando duas técnicas: tradução do corpus paralelo de português para galego utilizando o tradutor RBMT *Apertium* e um transliterador ortográfico para palavras desconhecidas. Este transliterador *ad hoc* converte a ortografia portuguesa para a castelhana, que é a utilizada atualmente em galego, com exceção de alguns grafemas (por exemplo, *x*, que é utilizado para *g*, *j* e *x*).

Passemos à análise dos resultados de um ponto de vista qualitativo e quantitativo. Do primeiro ponto de vista, vejamos como o *Carvalho SMT* traduz o seguinte texto da Wikipédia em inglês:

> Art is the process or product of deliberately arranging elements in a way that appeals to the senses or emotions. It encompasses a diverse range of human activities, creations, and modes of expression, including music, literature, film, sculpture, and paintings. The meaning of art is explored in a branch of philosophy known as aesthetics.

O resultado foi o seguinte:

> Arte é o proceso ou produto de arranxar deliberadamente elementos dunha forma que apela à sentidos ou emocións. Engloba un diversificado abano de actividades humanas, creacións e modos de expresión, inclusive da música, da literatura, filmes, escultura e pinturas. O significado de arte é explotada en un ramo da filosofía conhecida como aesthetics.

Veremos a seguir a mesma tradução feita pelo *Google Translate* nesse mesmo ano (2 de março de 2010):

> A arte é o proceso ou produto de deliberadamente organizar elementos de un modo que pide aos sentidos ou emocións. Engloba unha variada gama de actividades humanas, creacións, e modos de expresión, incluíndo a música, literatura, cine, escultura e pintura. O significado da arte é explotado desde unha rama da filosofía conhecido como estética.

De um ponto de vista quantitativo, para obter a métrica de qualidade de tradução automática BLEU, foi utilizado um pequeno corpus de referência padrão de 11 500 palavras, que foi desenvolvido na empresa imaxin|software como resultado da tradução manual de 500 frases de inglês para galego a partir da versão online do jornal inglês *The Guardian*. A Tabela 2 apresenta uma comparação entre os resultados em BLEU do *Carvalho SMT* e do *Google Translate* (desenvolvidos quase no mesmo período).

Na altura, este resultado indicava que um tradutor estatístico treinado a partir da transformação de corpora de português em galego era capaz de conseguir um tradutor de inglês-galego que, estando longe do *Google*

Translate, talvez pudesse, com a incorporação de mais corpora livres de inglês-português, conseguir melhores resultados na tradução automática entre inglês e galego.

Tabela 2. Medidas de comparação BLEU (em 2009) entre Google Translate e Carvalho SMT

SMT	**inglês-galego**	**galego-inglês**
Google Translate	25,59	35,91
Carvalho SMT	15,59	18,95

4.2 Tradutor neuronal de espanhol-galego e de inglês-galego com recursos do português

O ano 2021 viu nascer um projeto na Galiza com o nome *Proxecto Nós*, liderado pela Xunta de Galicia, que encomendou ao CITIUS e ao ILG o desenvolvimento de recursos, ferramentas e demonstradores inteligentes para galego. Um dos subprojectos é o desenvolvimento de sistemas de tradução automática neuronal (NMT), que arrancou no início de 2022.

Para seguirmos a metodologia, verificamos que há pouca distância linguística entre o português e o galego e notamos ainda que a Lei Paz-Andrade já foi aprovada em 2014. Também se deve ter em conta que, para o resultado de distância linguística, a ortografia desempenha um papel relevante (Pichel Campos, 2021), o que contribui para a possibilidade de utilização de um transliterador para a transformação de um corpus de português num corpus de galego. Uma vez que a Lei Paz-Andrade descongelou as relações linguísticas galego-portuguesas, o aumento do número de corpora para galego através da transformação de corpora de português tornou-se conveniente. Por todas estas razões, foi iniciado o desenvolvimento de um protótipo NMT de espanhol-galego com vista a gerar um Demonstrador final, sendo realizadas posteriormente diferentes experiências com arquiteturas *LSTM* e *Transformer* entre os pares espanhol-galego e inglês-galego, a partir da transliteração de corpora do português (Ortega et al., 2022).

Os corpora que utilizámos para a formação dos sistemas NMT de ES-GL e EN-GL foram recolhidos de Opus[16]. Utilizámos o *Europarl* para os pares ES-PT e EN-PT, com cerca de 2 milhões de frases por língua, e o *OpenSubtitles*, que contém mais de 30 milhões de frases em ES-PT e mais de 32 milhões em EN-PT. Além disso, adicionámos diretamente um minicorpus de espanhol-galego chamado *CLUVI* com 144 000 frases.

A nossa estratégia de treinar um tradutor automático neuronal de ES-GL e EN-GL baseado nos corpora de ES-PT e EN-PT transformados em corpora de ES-GL e EN-GL consistiu em duas etapas. Na primeira, utilizámos a transliteração para criar corpora paralelos de inglês-galego e de espanhol-galego com base na transliteração do português para o galego dos corpora de inglês-português e de espanhol-português, fazendo uso da ferramenta de transliteração *port2gal*[17]. Na segunda, com este corpus de galego (transliterado), treinámos duas arquiteturas *LSTM* e *Transformer modelo seq2seq* no sistema de tradução automática neural *OpenNMT*.

No que respeita à *LSTM*, as principais características foram: duas camadas ocultas, 500 unidades *LSTM* ocultas por camada, alimentação de entrada permitida, 13 *epochs*, *batch size* de 64. Alterámos também os parâmetros predefinidos do *learning rate* para 100 000 *training steps* e 10 000 *validation steps*. No caso do *Transformer*, descrito em Garg et al. (2019), este foi configurado com parâmetros de formação predefinidos: 6 camadas para codificação e descodificação e *batch size* de 4096 tokens. Também modificámos os parâmetros do *learning rate* para os mesmos valores que a configuração *LSTM*. Neste caso, utilizámos a tokenização das subpalavras, realizada com *SentencePiece*.

Os melhores resultados do nosso sistema para o par espanhol-galego numa arquitetura *LSTM* foram de 51,1 BLEU, que estão próximos do estado da arte em RBMT de espanhol-galego sem reutilização do português, que apresentam um BLEU acima de 60 (Bayón e Sánchez Gijón, 2019). Para o par inglês-galego os resultados do *BLEU*, resultantes da utilização de uma arquitetura *Transformer*, foram: 29,3, o que aumenta significativamente os resultados do *Carvalho SMT*, que apresenta um *BLEU* de 15,59. Um resumo final destes resultados encontra-se na Tabela 3:

16 https://opus.nlpl.eu/

17 https://gramatica.usc.es/~gamallo/port2gal.htm

Tabela 3. Medidas de comparação BLEU entre Apertium (RBMT), Carvalho (SMT) e Nós (NMT)

	espanhol-galego	**inglês-galego**
Apertium (RBMT)	> 60	-
Carvalho (SMT)	-	15,59
Nós (NMT)	51,1 (LSTM)	29,3 (Transformer)

Estes resultados mostram, por um lado, que a transformação de corpora de espanhol-português em espanhol-galego através da transliteração ortográfica gera resultados de tradução automática de espanhol-galego muito próximos do estado da arte na arquitetura *LSTM*. Por outro lado, mostra que, para o caso da tradução automática de inglês-galego, a mesma transformação, mas neste caso realizada no corpus de inglês-português, gera um tradutor melhor do que um tradutor estatístico com a arquitetura *Transformer*. Tal pode estar relacionado com uma melhor competência dos tradutores neuronais com arquitetura *Transformer*, com um corpus de inglês-português transliterado de maior dimensão do que no projeto *Carvalho*, ou com ambos os fatores.

5 Conclusões

Com base na ideia da relação controversa entre galego e português, criámos um método que pode ser aplicado a situações semelhantes (por exemplo, servo-croata, dividido em sérvio, bósnio e croata), com o objetivo de criar sistemas de tradução automática neuronal (NMT) para línguas com poucos recursos.

Uma vez que estes sistemas NMT necessitam fundamentalmente de grandes corpora de traduções humanas entre as línguas a traduzir (por exemplo, inglês-galego), e estes podem ser escassos para uma das línguas, utilizamos os corpora de tradução de uma língua próxima com suficientes recursos. Além disso, a fim de descobrir qual é esta língua vizinha, escolhemos uma métrica que calcula a distância entre corpora chamada *perplexity*, e observamos se existem quadros legislativos que permitam a utilização dos recursos desta língua próxima, para que o tradutor NMT a ser construído possa ser oficializado sem qualquer controvérsia.

No caso particular do galego, *perplexity* indicou que o galego, língua com poucos recursos, tem uma elevada proximidade com o português, que tem corpora de tradução suficientes em relação a outras línguas. Por outro lado, a aprovação da Lei Paz-Andrade para o uso do português e as relações com a Lusofonia abriram uma enorme via que permitiu o aproveitamento oficial do português para a criação de sistemas de tradução automática (NMT) para o galego.

Nos resultados obtidos no sistema NMT desenvolvido com corpora do português, tanto no par inglês-galego como no par espanhol-galego, observam-se resultados iniciais interessantes para o par inglês-galego, no caso dos sistemas SMT, e resultados próximos do estado da arte (RBMT) para o par espanhol-galego. Além disso, os sistemas de tradução neuronal inglês-galego que utilizam apenas a transliteração (*Nós NMT*) dos corpora inglês-português obtêm melhores resultados do que os sistemas de tradução estatística (*Carvalho SMT*) no mesmo par. Por fim, consideramos que estes resultados permitem considerar como desejável iniciar a produção de mais sistemas de tradução automática de alta qualidade (*NMT*) para o galego com base no aproveitamento do português.

6 Referências

Alegría Loinaz, Iñaki, Arantzabal, Iñaki, Forcada, Mikel, Gómez Guinovart, Xavier, Padró, Lluís, Pichel Campos, José Ramom, e Waliño, Josu. (2006). OpenTrad: Traducción automática de código abierto para las lenguas del Estado español. *Procesamiento del Lenguaje Natural*, *37*, 357–358.

Alonso Pintos, Serafín. (2005). Escribir e prescribir. Variación e prescrición ortográfica en lingua galega (1950–1982). Em R. Álvarez Blanco, e H. Monteagudo Romero (Eds.), *Norma lingüística e variación, unha perspectiva desde o idioma galego* (pp. 189–199). Consello da Cultura Galega e Instituto da Lingua Galega. https://doi.org/10.17075/nlv.2005.009

Álvarez Cáccamo, Celso, e Herrero Valeiro, Mário J. (1996). O "continuum" da escrita na Galiza: entre o espanhol e o português. *Agália: Publicaçom internacional da Associaçom Galega da Lingua*, *46*, 143–156.

Bahdanau, Dzmitry, Cho, Kyung Hyun, e Bengio, Yoshua. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. Em Yoshua Bengio, e Yann LeCun (Eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings* (pp. 1–15). https://doi.org/10.48550/arXiv.1409.0473

Basanta, Noemi, e Regueira, Xosé Luís. (2017, 28–30 de junho). *Este mapa talvez ajudava a romper algum preconceito: O valor múltiple da autenticidade no seo do capitalismo global* [Apresentação da conferência]. III Simposi Internacional EDiSo, Desigualtats i nous discursos socials, Universitat Pompeu Fabra, Barcelona, Spain.

Bayón, María do Campo, e Sánchez-Gijón, Pilar. (2019). Evaluating machine translation in a low-resource language combination: Spanish–Galician. Em Mikel Forcada, Andy Way, John Tinsley, Dimitar Shterionov, Celia Rico, e Federico Gaspari (Eds.), Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks (pp. 30–35). European Association for Machine Translation.

Berg, Ivar. (2016). The making of the Scandinavian languages. Em Gijsberg Rutten e Kristine Horner (Eds.), *Metalinguistic perspectives on Germanic languages: European case studies from past to present* (pp. 35–55). Peter Lang.

Busto Miramontes, Beatriz. (2016). *La Galicia proyectada por NO-DO. La arquitectura del estereotipo cultural a partir del uso del folclore musical (1943–1981)* [Tese de doutoramento]. Universidad Autónoma de Madrid.

Carvalhido, Xurxo. (2017). Revisão do livro *A imagem de Portugal na Galiza* de Carlos Quiroga. *Boletim da Academia Galega da Lingua Portuguesa*, *10*, 283–288.

Carvalho Calero, Ricardo. (1985). O problema ortográfico. *Agália: Publicaçom internacional da Associaçom Galega da Lingua*, *2*, 127–134.

Carvalho Calero, Ricardo. (1990). *Do galego e da Galiza*. Edicións Sotelo Blanco.

Chiswick, Barry R., e Miller, Paul W. (2005). *Journal of Multilingual and Multicultural Development*, *26*(1), 1–11. https://doi.org/10.1080/14790710508668395

Ciscel, Mathew H. (2006). A separate Moldovan language? *Nationalities Papers*, *34*(5), 575–597. https://doi.org/10.1080/00905990600952988

Climent Gallart, Jorge Antonio. (2020). El PIN parental y la jurisprudencia del TEDH. *Actualidad Jurídica Iberoamericana*, *13*, 102–121.

Comissão Europeia. (2011). Europe without borders. The Schengen area.

Conde, Valéria Gil. (2003). *História da normatização lingüística do galego: Tendências e conflitos* [Tese de doutoramento]. Universidade de São Paulo.

Constitución Española. (1978). *Boletín Oficial del Estado*, *311*.

Corcuera Atienza, Javier. (2000). La Constitución española de 1931 en la historia constitucional comparada. Em J. Varela Suanzes (Ed.), *Fundamentos: Cuadernos monográficos de teoría del estado, derecho público e historia constitucional*, *2*, 629–695.

Duarte Collazo, Silvia. (2014). O estándar galego: reintegracionismo vs. autonomismo. *Romanica Olomucensia*, *26*(1), 1–13. https://doi.org/10.5507/ro.2014.001

Fernández Souto, Ana Belén, Rúas Araújo, José, e Prada Iglesias, Adriana. (2019). Más allá de la proximidad cultural: vínculos entre el idioma gallego y el portugués en la negociación internacional. Em Silvia Montero Küpper, Montse Vázquez Gestal, e Iván Puentes Rivera (Eds.), *MonTI. Monografías de Traducción e Interpretación*, *5*, 5–71. https://doi.org/10.6035/MonTI.2019.ne5.2

Flor, Vicent. (2012). *« Llengua valenciana, mai catala »*. Sécessionnisme linguistique et revitalisation linguistique au Pays valencien (Espagne). *Lengas. Revue de sociolinguistique*, *72*, 133–151. https://doi.org/10.4000/lengas.119

Forcada, Mikel, Gisnestí-Rosell, Mireia, Nordfalk, Jacob, O'Regan, Jim, Ortiz Rojas, Sergio, Pére-Ortiz, Juan Antonio, Sánchez Martínez, Felipe, Ramírez-Sánchez, Gema, e Tyers, Francis. (2011). Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, *25*(2), 127–144. https://doi.org/10.1007/s10590-011-9090-0

Gamallo, Pablo, Pichel Campos, José Ramom, e Alegría, Iñaki. (2017). From language identification to language distance. *Physica A: Statistical Mechanics and its Applications*, *484*, 152–162. https://doi.org/10.1016/j.physa.2017.05.011

Garg, Sarthak, Peitz, Stephan, Nallasamy, Udhyakumar, e Paulik, Matthias. (2019). Jointly learning to align and translate with transformer models. Em Kentaro Inui, Jing Jiang, Vincent Ng, e Xiaojun Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP–IJCNLP)* (pp. 4453–4462). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1453

Geerts, Guido, Nootens, Johan, e van den Broeck, Jef. (1978). Flemish attitudes towards dialect and standard language: A public opinion poll. *International Journal of the Sociology of Language*, *15*, 33–46. https://doi.org/10.1515/ijsl.1978.15.33

Gómez Guinovart, Xavier. (1997). Traducción automática inglés-español: Estado del arte. Em Isabel Moskowich, Emma Lezcano e Santiago González Fernández-Corugedo (Eds.), *Some sundry wits gathered together* (pp. 31–40). Servizo de Publicacións da Universidade da Coruña.

Gómez Guinovart, Xavier. (2006). Tecnoloxías da lingua galega e normalización lingüística. Em Servizo de Normalización Linguística (Ed.), *Lingua e investigación. II Xornadas sobre lingua e usos* (pp. 79–91). Servizo de Publicacións da Universidade da Coruña.

González González, Manuel. (1985). La recuperación del gallego. *Revista de filología románica*, *3*, 101–120.

González González, Manuel. (2006). A investigación ao servizo da normalización da lingua galega na sociedade da información. Em Servizo de Normalización Linguística (Ed.), *Lingua e investigación. II Xornadas sobre lingua e usos* (pp. 147–163). Servizo de Publicacións da Universidade da Coruña.

Henríquez Salido, Maria do Carmo. (2020). O professor Manuel Rodrigues Lapa e a língua da Galiza. *Confluência: Revista do Instituto de Língua Portuguesa*, *58*, 9–32. https://doi.org/10.18364/rc.v1i58.373

Henríquez Salido, María do Carmo. (2021). O galego-português, matriz do mundo lingüístico luso-brasileiro. *Caplletra. Revista Internacional de Filologia*, *71*, 57–94. https://doi.org/10.7203/caplletra.71.21032

Herrero Valeiro, Mário J. (2003a). Ilusões glotopolíticas e planificação linguística na Galiza. Em María del Carmen Cabeza Pereiro, Anxo M. Lorenzo Suárez, e Xoán Paulo Rodríguez Yáñez (Eds.), *Comunidades e individuos bilingües. Actas do I Simposio Internacional sobre o Bilingüismo: Universidade de Vigo, Galicia–Spain, 21–25 outubro–october 1997* (pp. 1058–1069). Universidade de Vigo, Servizo de Publicacións.

Herrero Valeiro, Mário J. (2003b). The discourse of language in Galiza: Normalisation, diglossia, and conflict. *Sociolinguistic Studies*, *3–4*, 289–320. https://doi.org/10.1558/sols.v4i1.289

Herrero Valeiro, Mário J. (2011). *Guerra de grafias e conflito de elites na Galiza contemporânea*. Através.

Instituto da Lingua Galega e Real Academia Galega. (2004). *Normas ortográficas e morfolóxicas do idioma galego*. Galaxia.

Isphording, Ingo Eeduard, e Otten, Sebastian. (2013). The Costs of Babylon-Linguistic Distance in Applied Economics. *Review of International Economics*, *21*(2), 354–369. https://doi.org/10.1111/roie.12041

Jauhiainen, Tommi, Lui, Marco, Zampieri, Marcos, Baldwin, Timothy, e Lindén, Krister. (2019). Automatic language identification in texts: A survey. *Journal of Artificial Intelligence Research*, *65*, 675–782. https://doi.org/10.1613/jair.1.11675

Koehn, Philipp. (2005). Europarl: A parallel corpus for statistical machine translation. *Proceedings of Machine Translation Summit X: Papers* (pp. 79–86).

Koehn, Philipp. (2009). *Statistical machine translation*. Cambridge University Press. https://doi.org/10.1017/CBO9780511815829

Koehn, Philipp. (2020). *Neural machine translation*. Cambridge University Press. https://doi.org/10.1017/9781108608480

Koehn, Philipp, Hoang, Hieu, Birch, Alexandra, Callison-Burch, Chris, Federico, Marcello, Bertoldi, Nicola, Cowan, Brooke, Shen, Wade, Moran, Christine, Zens, Richard, Dyer, Chris, Bojar, Ondřej, Constantin, Alexandra, e Herbst, Evan. (2007). Moses: Open-source toolkit for statistical machine translation. Em Sophia Ananiadou (Ed.), *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics. Companion volume. Proceedings of the demo and poster sessions* (pp. 177–180). Association for Computational Linguistics https://doi.org/10.3115/1557769.1557821

Labraña, Sabela, e Vázquez, Ignacio. (1997). Escribir en una lengua que escuchamos robada. *Quimera. Revista de Literatura*, *158–159*, 91–94.

Lai, Mirko, Patti, Viviana, Ruffo, Giancarlo, e Rosso, Paolo. (2018). Stance evolution and Twitter interactions in an Italian political debate. Em Max Silberztein, Faten Atigui, Elena Kornyshova, Elisabeth Métais, e Farid Meziane (Eds.), *International Conference on Applications of Natural Language to Information Systems* (pp. 15–27). Springer. https://doi.org/10.1007/978-3-319-91947-8_2

Lastra Muruais, Xosé. (2010). O galego e a educación. *Galegos= Gallegos*, *9*, 127–131.

Gobierno de España. (1981). Lei Orgánica 1/1981, do 6 de abril, do Estatuto de Autonomía para Galicia. *Boletín Oficial del Estado: martes 28 de abril de 1981, Núm. 101*, 8997–9003.

Lui, Marco, e Cook, Paul. (2013). Classifying English documents by national dialect. Em Sarvnaz Karimi e Karin Verspoor (Eds.), *Proceedings of the Australasian Language Technology Association Workshop* (pp. 5–15). Australasian Language Technology Association.

Maia, Clarinda de Azevedo. (1987). *História do galego-português: Estado linguístico da Galiza e do noroeste de Portugal desde o século XIII ao século XVI: com referência á situação do galego moderno* [Tese de doutoramento]. Universidade de Coimbra.

Malvar Fernández, Paulo, Pichel Campos, José Ramom, Senra, Óscar, e García, Alberto. (2010). Vencendo a escassez de recursos computacionais. Carvalho: Tradutor Automático Estatístico Inglês-Galego a partir do corpus paralelo Europarl Inglês-Português. *Linguamática*, *2*(2), 31–38.

Mariño Paz, Ramón. (2002). A obra lingüística de Carvalho Calero. Em Teresa Lopez e Francisco Salinas (Eds.), *Ricardo Carvalho Calero, Memoria do Século* (pp. 67–106). Universidade da Coruña. https://doi.org/10.17979/spudc.9788497497671.067

Mariño Paz, Ramón. (2008). *Historia de la lengua gallega*. Lincom Europa.

Mattoso, José. (1986). *A formação da nacionalidade*. Imprensa Nacional–Casa da Moeda.

Mira, Jorge, e Paredes, Ángel. (2005). Interlinguistic similarity and language death dynamics. *Europhysics Letters*, *69*(6), 1031–1034.

Mira Pérez, Jorge, e Paredes, Ángel. (2007). Modelado de la competición lingüística gallego-castellano: definición y obtención de distancia interlingüística y extrapolación a futuro. *Revista Real Academia Galega de Ciencias*, *26*, 5–16. https://doi.org/10.1209/epl/i2004-10438-4

Miranda Gonçalves, Rubén. (2016). La iniciativa legislativa popular como mecanismo de democracia participativa: Especial referencia a la Comunidad Autónoma de Galicia. Em Rubén Miranda Gonçalves (Ed.), *Administración pública, juventud y democracia participativa* (pp. 153–164). Dirección Xeral de Xuventude e Voluntariado.

Mobergy, Jens, Gooskens, Charlotte, Nerbonney, John, e Vaillette, Nathan. (2007). Conditional entropy measures intelligibility among related languages. *LOT Occasional Series*, *7*, 51–66.

Molina, Giovanni, AlGhamdi, Fahad, Ghoneim, Mahmoud, Hawwari, Abdelati, Rey-Villamizar, Nicolas, Diab, Mona, e Solorio, Thamar. (2016). Overview for the second shared task on language identification in code-switched data. Em Mona Diab, Pascale Fung, Mahmoud Ghoneim, Julia Hirschberg, e Thamar Solorio (Eds.), *Proceedings of the Second Workshop on Computational Approaches to Code Switching* (pp. 40–49). Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-5805

Monteagudo, Henrique. (1990). Sobre a polémica da normativa do galego. *Grial*, *28*(107), 294–316.

Monteagudo, Henrique. (2017). A lingua no tempo, os tempos da lingua. O galego, entre o portugués e o castelán. Em Marta Negro Romero, Rosario Álvarez Blanco, e Eduardo Moscoso Mato (Eds.), *Gallæcia. Estudos de linguistica portuguesa e galega* (pp. 1760). Universidade de Santiago de Compostela.

Monteagudo, Xoaquín. (1990). A oficialidade da lingua galega: O marco legal. *Grial*, *28*(107), 387–398.

Montero Santalha, José-Martinho. (2001). Alguns testemunhos de reintegracionismo lingüístico galego-português nos anos 60–70. *Agália: Publicaçom Internacional da Associaçom Galega da Lingua*, *65*, 9–16.

Montero Santalha, José-Martinho. (2008). O nome da Galiza. *Boletim da Academia Galega da Língua Portuguesa*, *1*, 11–34.

Montero Santalha, José-Martinho. (2010). A lusofonia e a língua portuguesa da Galiza: Dificuldades do presente e tarefas para o futuro. Em Comissom Lingüística da AGAL (Ed.), *Por um galego extenso e útil: Leituras da língua de aquém e de além* (pp. 35–55). Através.

Morell, Xosé Carlos. (2019). A Lei Paz Andrade: O valor da categoria internacional. *Nós no mundo: A lusofonía en Galicia, entre a Lei Paz Andrade e a EGAEX*, *10*, 6–7.

Mønnesland, Svein. (1997). Emerging literary standards and nationalism. The disintegration of Serbo-Croatian. Em María del Carmen Cabeza Pereiro, Anxo M. Lorenzo Suárez, e Xoán Paulo Rodríguez Yáñez (Eds.), *Actas do I Simposio Internacional sobre o Bilingüismo* (pp. 1103–1113). Universidade de Vigo, Servizo de Publicacións

Muñoz Carrobles, Diego. (2008). Evolución e cambios na normativa oficial do galego (1982–2007). *Madrygal: Revista de Estudios Gallegos*, *11*, 49–56.

Nerbonne, John, e Heeringa, Wilbert (1997). Measuring dialect distance phonetically. *Proceedings of the Third Meeting of the ACL Special Interest Group in Computational Phonology* (pp. 11–18).

Nogueira López, Alba. (2020). Crónica Lexislativa de Galicia 2019: Fin de lexislatura coa normalización lingüística baixo mínimos. *Revista de Llengua i Dret*, *73*, 216–220.

Ochoa Monzó, Josep. (2006). Antonio Xavier Ferreira Fernández; Alba Nogueira López; Anxo Tato Plaza, Luis Villares Naviera, «Estatuto xurídico da lingua galega». *Revista de Llengua i Dret*, *45, 345-349*.

Ortega, John E., Castro Mamani, Richard, e Cho, Kyunghyun. (2020). Neural machine translation with a polysynthetic low resource language. *Machine Translation*, *34*(4), 325–346. https://doi.org/10.1007/s10590-020-09255-9

Ortega, John E., de-Dios-Flores, Iria, Gamallo, Pablo, e Pichel, José Ramom. (2022). A neural machine translation system for Galician from transliterated Portuguese text. Em Miguel Alonso, Margarita Alonso-Ramos, Carlos Gómez-Rodríguez, David Vilares, e Jesús Vilares (Eds.), *Proceedings of the Annual Conference of the Spanish Association for Natural Language Processing. CEUR Workshop Proceedings vol. 3224* (pp. 92–95).

Pacho Blanco, José Manuel. (2016). *Lenguas y Constitución: El artículo 3 de la Constitución española* [Tese de doutoramento]. Universidade de Vigo.

Pérez-Barreiro Nolla, Fernando. (1990). Which language for Galicia? The status of Galician as an official language and the prospects for its reintegration with Portuguese. *Portuguese Studies*, *6*, 191–210.

Pérez González, Manuel, e Martínez Puñal, Antonio (1986). *El Estatuto de Autonomía de Galicia como elemento de institucionalización de las relaciones culturales galaico-portuguesas*. Universidad del País Vasco.

Petroni, Filippo, e Serva, Maurizio. (2010). Measures of lexical distance between languages. *Physica A: Statistical Mechanics and its Applications*, *389*(11), 2280–2283. https://doi.org/10.1016/j.physa.2010.02.004

Pichel Campos, José Ramom. (2010). *Google é reintegracionista*. A Nossa Galáxia.

Pichel Campos, José Ramom. (2020). *Medidas de distância entre línguas baseadas em corpus. Aplicação à linguística histórica do galego, português, espanhol e inglês* [Tese de doutoramento]. Euskal Herriko Unibertsitatea/Universidad del País Vasco.

Pichel, José Ramom, Gamallo, Pablo, Alegría, Iñaki, e Neves, Marco. (2021). A methodology to measure the diachronic language distance between three languages based on perplexity. *Journal of Quantitative Linguistics*, *28*(4), 306–336. https://doi.org/10.1080/09296174.2020.1732177

Pichel Campos, José Ramom, Malvar, Paulo, Senra Gómez, Oscar, Gamallo, Pablo, e García, Alberto. (2009). Carvalho: Un sistema de traducción estadística inglés-galego construído a partir del corpus paralelo inglés-portugués EuroParl. *Procesamiento del Lenguaje Natural*, *43*, 379–381.

Pichel Campos, José Ramom, Vázquez, Diego, Castro, Luz, e Fernández, Antonio. (2014). imaxin|software: PLN aplicada a la mejora de la comunicación multilingüe de empresas e instituciones. *Procesamiento del Lenguaje Natural*, *53*, 189–192.

Piller, Ingrid. (2016). *Linguistic diversity and social justice: An introduction to applied sociolinguistics*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199937240.001.0001

Quiroga, Carlos. (2013). Tradução e Jogo de Tronos-entre Culturas de Falares polissimétricos. Em Mosquera Carregal e Xesús Manuel (Eds.), *Lingua e tradución: IX Xornadas sobre Lingua e Usos* (pp. 203–215). Servizo de Publicacións da Universidade de Santiago de Compostela.

Ramallo, Fernando. (2016). *As linguas minorizadas no Estado español* (Informe presentado no 82). Congreso do PEN Internacional.

Ramallo, Fernando, e Rei-Doval, Gabriel. (2015). The standardization of Galician. *Sociolinguistica*, *29*(1), 61–82. https://doi.org/10.1515/soci-2015-0006

Regueiro Tenreiro, Manuel. (1994). Potenciación del uso del gallego y protección de la lengua. *Euskera*, *1*, 39–44.

Rodrigues Fagim, Valentim. (2001). *O galego (im)possível*. Através.

Rodríguez Fernández, Elías. (2018). *Análise e balanço do quadro legislativo para o relacionamento intercomunitário: O caso da Lei Paz-Andrade* [Traballo fin de Grao]. Universidade da Coruña.

Rodríguez González, María del Mar, Marauri Castillo, Iñigo, e Pérez Dasilva, Jesús Ángel (2016). Itzultzaile automatikoak euskaraz, bilakaera eta prestazioak. *Mediatika. Cuadernos de Medios de Comunicación*, *15*, 225–242.

Salgado, Benigno F., e Monteagudo, Henrique. (1993). The standardization of Galician: The state of the art. *Portuguese Studies*, *9*, 200–213.

Samartim, Roberto. (2012). Língua somos: A construção da ideia de língua e da identidade coletiva na Galiza (pré-) constitucional. Em Olivia Rodríguez-González, Laura Carballo Piñeiro, e Burghard Baltrusch (Eds.), *Novas achegas ao estudo da cultura galega II: Enfoques socio-históricos e lingüístico-literarios* (pp. 27–36). Universidade da Coruña.

Sánchez Vidal, Paulo. (2005). Unha achega ó estudo do proceso de codificación ortográfica e gramatical da lingua galega (1980–2000). Em Rosario Álvarez Blanco e Henrique Monteagudo Romero (Eds.), *Norma lingüística e variación, unha perspectiva desde o idioma galego* (pp. 201–221). Consello da Cultura Galega. https://doi.org/10.17075/nlv.2005.010

Silva Domínguez, Carme. (2008). Notas sobre "anceio", un probable lusismo no galego de preguerra. Em Mercedes Brea López, Francisco Fernández Rei, e Xosé Luis Regueira Fernández (Eds.), *Cada palabra pesaba, cada palabra medía: Homenaxe a Antón Santamaría* (pp. 181–192). Universidade de Santiago de Compostela.

Teyssier, Paul. (1982). *História da língua portuguesa*. Martins Fontes.

Torres Feijó, Elías J. (2000). Norma lingüística e (inter-)sistema cultural: O caso galego. Em Juan María Carrasco González, Maria Luísa Leal, e María Jesús Fernández García (Eds.), *1er Encuentro Internacional de lusitanistas españoles* (pp. 967–998). Universidad de Extremadura.

Torres Feijó, Elías J. (2019). Cultura portuguesa e legitimação do sistema galeguista: Historiadores e filólogos (1880–1891). Em Roberto Samartim e Carlos Pazos Justo (Eds.), *Portugal e(m) nós: Contributos para a compreensão do relacionamento cultural galego-português* (pp. 41–84). Húmus.

Venâncio, Fernando. (2020). *Assim nasceu uma língua: Sobre as origens do português*. Guerra e Paz Editores.

Villares Naveira, Luis. (2010). Estudo sobre a legalidade do "Decreto 79/2010 para o plurilingüismo no ensino non universitario de Galicia": Unha perspectiva constitucional. Em Luis Villares Naveira (Ed.), *Estudos xurídicos sobre o decreto para o plurilingüismo* (pp. 9–56). Laiovento.

Xunta de Galicia. (2010). Decreto 79/2010, do 20 de maio, para o plurilingüismo no ensino non universitario de Galicia. *Diario Oficial de Galicia*, *97*, 9242–9247.

Xunta de Galicia. (2014). Lei 1/2014, do 24 de marzo, para o aproveitamento da lingua portuguesa e vínculos coa lusofonía (Lei Paz-Andrade). *Diario Oficial de Galicia*, *68*, 15608–15610.

Xunta de Galicia. (1983). Lei 3/1983, do 15 de xuño, de normalización lingüística. *Diario Oficial de Galicia*, *84*, 1893–1899.

Zampieri, Marcos, Gebre, Binyam Gebrekidan, Costa, Hernani, e van Genabith, Josef. (2015). Comparing approaches to the identification of similar languages. Em Preslav Nakov, Marcos Zampieri, Petya Osenova, Liling Tan, Cristina Vertan, Nikola Ljubešić, e Jörg Tiedemann (Eds.), *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects* (pp. 66–72). Association for Computational Linguistics.