Contributos para o Aumento de Qualidade na Língua Digital

Anabela Barreiro
INESC-ID
abarreiro@inesc-id.pt

Abstract

As tecnologias da língua entraram na esfera da vida quotidiana do cidadão comum em diversas áreas: na comunicação social e nos media, no ensino e aprendizagem da língua e de línguas estrangeiras, em divulgação cultural, na saúde, nas relações interpessoais nas redes sociais e em relações internacionais, entre outros. Este processo de globalização linguística representa simultaneamente um desafio e uma oportunidade para os falantes do português. Este artigo descreve algumas ferramentas e recursos linguísticos desenvolvidos no INESC-ID que visam um aumento da qualidade do português em aplicações de linguagem natural: o eSPERTo, o CLUE-Aligner, o e-PACT, e as CLUE Alignment Guidelines. O eSPERTo é uma plataforma de geração de paráfrases com vista à reescrita de textos e ferramenta para o auxílio à tradução. O CLUE-Aligner é um alinhador interativo que permite a anotação e extração de unidades lexicais multipalavra e outras unidades frásicas em corpos paralelos de traduções e em textos parafrásticos. O e-PACT é um corpo de paráfrases de unidades lexicais multipalavra e expressões alinhadas construído a partir de traduções do inglês para as variantes europeia e brasileira do português. As CLUE Alignment Guidelines são linhas diretrizes de alinhamento de paráfrases e unidades de tradução desenvolvidas a partir de corpos paralelos das variantes europeia e brasileira do português e de corpos paralelos de inglês e de línguas românicas. Para além de visarem a qualidade linguística, estas ferramentas estão também direcionadas para a internacionalização da língua portuguesa.

1 Introdução

Apesar dos avanços significativos sobretudo na última década nas tecnologias da língua, ainda se registam lacunas ao nível da qualidade na língua digital[1], área onde

[1]Língua digital é toda a língua que passa por computadores e ferramentas eletrónicas.

não tem sido feita uma aposta suficientemente forte e concreta à escala europeia ou global. A garantia ou controlo de qualidade linguística é uma técnica que garante a correção e integridade linguística de um sistema e deveria representar uma etapa crucial em projetos de linguagem natural, não apenas devido às exigências dos utilizadores no que respeita a questões linguísticas, mas também porque os erros que os sistemas apresentam podem influenciar o comportamento e a confiança dos utilizadores em relação à tecnologia e ao seu potencial.

A ausência de um plano de ação estratégico apropriado para lidar com lacunas linguísticas motiva, por um lado, a criação de novas metodologias de avaliação (Barreiro et al. 2013, 2014a) e, por outro, a criação das novas ferramentas e recursos linguísticos, tais como as que descrevemos resumidamente neste artigo. Mais do que apresentar uma solução imediata para a utilização profissional das ferramentas que desenvolvemos, o nosso contributo tem o ensejo de reafirmar a necessidade de mais investigação nesta área que tenha como principais objetivos (i) o aumento significativo da qualidade linguística das aplicações em benefício da sociedade em geral e (ii) o aumento do potencial de crescimento da tecnologia, tanto para as aplicações mais simples, como para as que envolvem um processo de globalização ou internacionalização das línguas, incluindo, de um modo especial, o português, para o qual já existe a proposta de criação do português internacional (Santos 2014, 2015).

Apresentamos aqui duas ferramentas da tecnologia da língua e dois recursos linguísticos em desenvolvimento no Laboratório de Sistemas de Língua Falada (L2F) do INESC-ID. As ferramentas eSPERTo e CLUE-Aligner visam a sua aplicabilidade imediata nos domínios da comunicação e didática da língua, que vão desde a escrita de textos ao enriquecimento do conhecimento linguístico de um agente virtual inteligente e à sumarização. Os recursos linguísticos e-PACT e CLUE Alignment Guidelines contemplam, a um prazo mais alargado, a tradução automática de língua escrita, uma aplicação de grande impacto na globalização e internacionalização da língua portuguesa. Neste domínio, descrevemos os trabalhos em desenvolvimento, nomeadamente a tarefa de alinhamento de unidades multilingues para a tradução automática contemplando a integração de conhecimento semântico-sintático e parafrástico em sistemas de tradução automática.

2 eSPERTo: Sistema de Parafraseamento

O projeto eSPERTo[2] visa o desenvolvimento de um sistema parafraseador com o mesmo nome – eSPERTo é um acrónimo para **S**istema de **P**arafraseamento para

[2]O eSPERTo foi financiado pela Fundação para a Ciência e a Tecnologia, um projeto exploratório com a referência EXPL/MHC-LIN/2260/2013.

Edição e **R**evisão de **T**extos[3] – que integra conhecimento semântico e gramatical e tem em consideração o contexto. O sistema híbrido comporta um módulo de geração de paráfrases com base em gramáticas locais, um módulo de aquisição parafrástica baseado em estatística e em alinhamentos de unidades lexicais multipalavra e outras construções parafrásticas, e finalmente, uma aplicação interativa na Web.[4] Em linhas gerais, o termo paráfrase refere-se à relação entre duas ou mais construções relacionadas morfossintática e semanticamente. Consideramos que o nível de equivalência parafrástica pode ir desde uma unidade lexical multipalavra, como uma construção com um verbo suporte (*ter febre* = *estar febril*) até um contexto mais alargado, como uma frase (*podemos, logo à primeira vista, detetar uma diferença* (voz ativa) = *uma dessas diferenças pode ser observada de relance* (voz passiva)). Em muitos casos, uma relação parafrástica estabelece-se entre construções que correspondem à mesma unidade sintática, mas pode estabelecer-se também entre unidades sintáticas distintas numa relação sintático-semântica entre duas ou mais frases e/ou entre os seus constituintes.

O eSPERTo foi desenvolvido com base nos recursos bilingues do sistema OpenLogos (Barreiro et al. 2011, 2014b), uma versão em código aberto do sistema comercial Logos (Scott 2003; Scott 1989), recursos esses que posteriormente foram adaptados e integrados no ambiente linguístico NooJ (*La formalisation des langues: l'approche NooJ*; Silberztein 2016). A integração no módulo do português, ou seja, um conjunto de recursos chamado Port4NooJ, presentemente na sua terceira versão (Mota, Carvalho, and Barreiro 2016), envolveu a inclusão de relações de derivação e relações morfossintáticas e semânticas específicas da língua portuguesa. O sistema de parafraseamento reconhece unidades lexicais multipalavra e outras unidades frásicas com conhecimento semântico-sintático, que constitui o "motor" do sistema, e transforma-as em unidades ou expressões semanticamente idênticas ou equivalentes. As equivalências semânticas, neste caso, os pares de unidades parafrásticas, têm utilidade em muitas aplicações distintas. Um parafraseador inteligente, que tem por base conhecimento semântico, pode ser usado em sistemas de edição e revisão de textos, em sistemas de ensino da língua, em sistemas de resposta a perguntas, em sumarização, em sistemas de pesquisa para otimizar a busca de documentos, em sistemas de tradução, entre outros. De momento, a utilidade das capacidades parafrásticas do eSPERTo está a ser testada em dois cenários: (i) num sistema de resposta a perguntas para enriquecer a base de conhecimento e aumentar a qualidade linguística da interação conversacional entre um agente virtual inteligente e o seu utilizador e (ii) num sistema de sumarização para pré-processar

[3] O acrónimo do eSPERTo também funciona em inglês – **S**ystem for **P**araphrasing in **E**diting and **R**evision of **T**exts.

[4] O sistema parafrástico eSPERTo está disponível na internet a partir de: `https://esperto.l2f.inesc-id.pt/esperto/esperto/demo.pl`

textos, permitindo reduzi-los ou normalizá-los de modo a torná-los mais pequenos e/ou sistemáticos. Futuramente, está prevista a aplicação de unidades parafrásticas em tradução automática. As secções 2.1–2.3 ilustram as tarefas que parafraseamento nas aplicações supramencionadas. Atualmente, estamos a usar um sistema de processamento de linguagem natural desenvolvido no L2F, a STRING (Baptista, Mamede, and Markov 2014), para realizar tarefas de desambiguação, análise sintática, reconhecimento de entidades nomeadas, entre outras. As paráfrases geradas automaticamente pelo eSPERTo são posteriormente validadas por linguistas antes da sua integração no Port4NooJ.

2.1 Edição e Revisão de Textos

O parafraseamento é usado como técnica para articular uma ideia ou veicular informação de forma alternativa, facilitando a sua compreensão e tornando os textos mais valiosos para a sua audiência. As alternativas parafrásticas aumentam o poder expressivo dos sistemas de geração de linguagem natural ao permitirem a produção e pré-edição de textos mais variados e fluentes, ao mesmo tempo que tornam possível realizar um controlo da qualidade linguística, através de escolhas de desambiguação, eliminação de redundâncias, simplificação e uniformização da produção de frases. De acordo com a função e propósito do seu texto, o utilizador pode optar por paráfrases com um menor número de palavras, o que permite encurtar um texto[5] ou optar por alternativas parafrásticas que melhorem o significado das frases. Para além disso, as paráfrases ajudam a moldar o texto ao gosto e estilo do utilizador, que pode optar por uma ou outra expressão de acordo com a sua preferência estilística.

A aplicação mais generalizada do eSPERTo, mas também aquela que exige um maior rigor na utilização das unidades parafrásticas em contexto, é a da edição e revisão de textos. Esta aplicação, ainda em fase inicial de exploração, tem o objetivo de auxiliar editores, revisores e tradutores na preparação dos seus textos até se tornarem um produto final. As várias funcionalidades da atual interface Web do eSPERTo estão representadas na Figura 1.[6] Esta interface inclui mecanismos de edição de texto que permitem escolher entre uma variedade de alternativas para cada unidade lexical multipalavra ou expressão. O utilizador da plataforma parafrástica opta por uma das sugestões apresentadas pelo sistema e

[5]Para aplicações como a tradução automática, a redução do número de palavras de um texto é uma técnica de pré-edição comum que visa aumentar a qualidade da tradução desse mesmo texto.

[6]As versões mais antigas da arquitetura do eSPERTo foram disponibilizadas sob os nomes de Reescreve, uma versão para a língua portuguesa apresentada em Barreiro (2009) e de Rewriter (também com o nome de SPIDER), uma versão para a língua inglesa publicada em Barreiro and Cabral (2009) e Barreiro (2011).

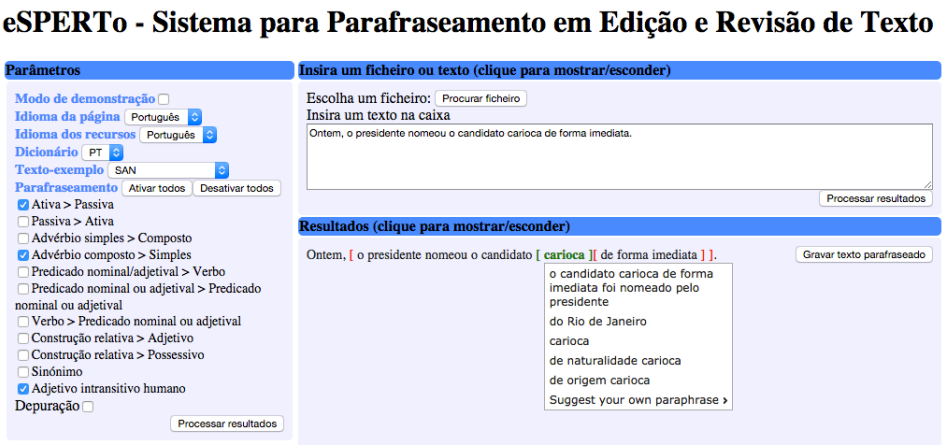

Figure 1: Uso interativo do eSPERTo na edição e revisão de textos

vê essa opção imediatamente aplicada ao seu texto. As sequências que o utilizador pode parafrasear (ou seja, as sequências identificadas pelo sistema como parafraseáveis) são apresentadas entre parênteses retos ([]). Por exemplo, o adjetivo intransitivo *carioca* pode ser parafraseado por expressões equivalentes, como *do Rio de Janeiro*, *de naturalidade carioca* ou *de origem carioca*, entre outras. No exemplo, a frase ativa também pode ser parafraseada por uma frase passiva equivalente. As paráfrases correspondentes geradas através da invocação do motor linguístico do NooJ são apresentadas na lista da caixa suspensa. Quando o utilizador move o cursor sobre a lista pode selecionar qualquer uma das sugestões parafrásticas. A sugestão selecionada torna-se mais destacada. O utilizador pode também inserir a sua própria sugestão. As sugestões inseridas pelos utilizadores serão validadas antes de serem adicionadas aos recursos linguísticos. Para além disso, o utilizador pode escolher alguns dos parâmetros da ferramenta, como a língua dos textos, os diferentes tipos de construção a parafrasear, podendo também escolher um texto-exemplo, carregar o seu próprio texto ou escrever diretamente na caixa de texto. Ao selecionar os tipos de construção, o utilizador está a selecionar os recursos que deseja aplicar mesmo sem ter um controlo direto sobre as gramáticas que serão usadas para processar o seu texto. Assim, as gramáticas de geração de paráfrases não são selecionáveis, apenas o tipo de transformação o é (por exemplo, a conversão de um predicado nominal num verbo ou um advérbio composto num advérbio simples e vice-versa). São as construções envolvidas nos processos transformacionais que são apresentadas ao utilizador como sugestões de melhoria da qualidade do seu texto.

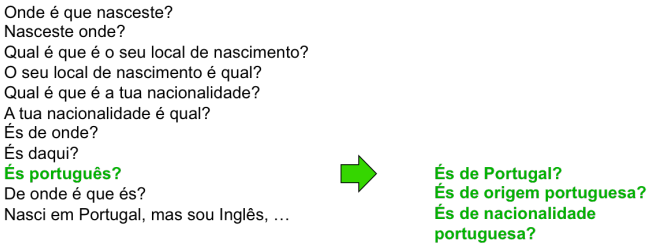

Figure 2: Paráfrases para a pergunta *És português?*

2.2 Sistemas de Resposta a Perguntas

As paráfrases são importante nos sistemas de resposta a perguntas, nomeadamente na validação de respostas, isto é, na indicação de que várias respostas equivalentes estão corretas em relação a determinada pergunta. Neste momento, a utilidade das paráfrases geradas pelo eSPERTo está a ser explorada e testada num agente conversacional inteligente que responde a perguntas feitas pelos seus utilizadores. Um sistema inteligente de parafraseamento que utiliza informação semântica é imprescindível para melhorar a capacidade de interação do agente com o utilizador humano. Os recursos parafrásticos permitem ao agente conversar de forma cada vez mais natural, respondendo a perguntas às quais, sem esse conhecimento, não conseguiria responder. Até ao momento, a tecnologia utilizada pelos agentes conversacionais é baseada em modelos simples essencialmente com recurso à estatística. Esses modelos permitem calcular ou medir a distância lexical das palavras de uma frase pronunciada por um utilizador e as perguntas que estão armazenadas na base de conhecimento do agente. Recorrendo ao modelo da proximidade lexical, a pergunta que tiver a distância mais curta em relação à frase do utilizador, acionará ou desencadeará a atribuição da resposta. As paráfrases geradas pelo eSPERTo enriquecerão a base de conhecimento do agente fornecendo uma ligação morfossintática e semântica entre perguntas ou entre respostas equivalentes, tal como está ilustrado na Figura 2. As paráfrases representadas no exemplo resultam de um processo de transformação de um adjetivo intransitivo humano de nacionalidade, *português*, nos seus equivalentes parafrásticos: *de Portugal*, *de origem portuguesa*, *de nacionalidade portuguesa*. A estas sugestões ter-se-ão de incluir as alternativas normalmente usadas pelos falantes de português do Brasil, *Você é português?*, com todas as suas variações.

Figure 3: Aplicação de paráfrases em sumarização

2.3 Sumarização

No caso da sumarização, as paráfrases permitem que a informação em vários documentos seja condensada e optimizada, ajudando simultaneamente a melhorar a qualidade linguística dos resumos. Um sistema inteligente de parafraseamento utiliza informação semântica para pré-processar textos antes de tarefas de sumarização. As paráfrases do eSPERTo estão também a ser exploradas para pré-processar texto no sentido de melhorar a deteção de conteúdo importante em tarefas de simplificação e de compactação. Neste contexto, o parafraseamento envolve a redução de texto e a uniformização lexical que servem de base para melhorar a legibilidade do texto, embora com alterações ao nível estilístico. A Figura 2.3 ilustra a utilização de unidades parafrásticas na tarefa de sumarização onde estão envolvidas (i) a transformação de um sintagma preposicional (*do passado*) num particípio passado (*passadas*), (ii) a transformação de uma construção com verbo suporte aspetual (*engajar-se na luta*) num verbo simples que lhe está morfossintática e semanticamente relacionado (*lutar*), (iii) a transformação de uma expressão verbal com nome predicativo (*lutar pela aprovação*) noutra expressão equivalente com predicado verbal (*lutar para aprovar*) e (iv) a transformação de um sintagma preposicional com a função de adjunto adnominal (*na Europa*) num adjetivo (*europeia*).

3 e-PACT – Corpo de Traduções EN-PE/PB

O e-PACT (**e**SPERTo **P**araphrase **A**ligned **C**orpus of **T**ranslations) (Barreiro and Mota 2016) é um corpo de paráfrases alinhadas que consiste em traduções de dois livros ingleses do autor David Lodge para as variantes europeia (PE) e brasileira (PB) do português, extraídas do COMPARA[7] (Frankenberg-Garcia and Santos 2003). O nosso objetivo não é o de "traduzir" entre variantes, porque o conceito de tradução não se aplica dentro de uma mesma língua, que carateriza coletivamente falantes que se entendem e que partilham "mundos" inseparáveis. O nosso objetivo é antes permitir e facilitar a acomodação, adaptação e intervariação dentro da mesma língua, de modo a enriquecer as opções linguísticas dos seus falantes inde-

[7] O COMPARA é um corpo paralelo bidirecional contendo 32 obras em inglês e 40 obras em português, onde cada frase inglesa está alinhada com a sua tradução em português e vice-versa.

pendentemente do lado do oceano em que se encontrem (ou dos oceanos, quando estendermos o nosso trabalho a outras variantes do português).

Para a construção do e-PACT, alinharam-se unidades lexicais multipalavra, expressões e outras unidades parafrásticas semanticamente equivalentes entre as frases das duas variantes de português. Estas frases foram previamente recolhidas através da consulta do COMPARA para alinhamentos contendo, na tradução em português desses textos, palavras que potencialmente pudessem desencadear as gramáticas de parafraseamento do eSPERTo (por exemplo, gramáticas de parafraseamento de adjetivos intransitivos humanos, construções com verbos suporte, entre outras). Apesar de o eSPERTo ter como alvo um conhecimento parafrástico independentemente da variante de português, escolhemos traduções em PE e PB do mesmo livro em inglês por motivos diferentes: (i) por constituírem corpos paralelos/comparáveis já alinhados à frase, (ii) por serem uma fonte rica em paráfrases do português "global", (iii) porque queríamos testar a nossa hipótese de que uma paráfrase em PE nem sempre é usada (ou tão comummente usada) em PB e vice-versa. As unidades parafrásticas recolhidas contêm expressões que são usadas exclusivamente por falantes de uma das variantes e expressões que são usadas em comum (a grande maioria) pelos falantes do português. Uma maior quantidade de dados permitirá a escolha ou adaptação de paráfrases para cada variante do português, mas apenas por opção. Neste sentido, o utilizador do eSPERTo poderá escolher, de entre todas as paráfrases sugeridas, as que possam ser mais naturais na sua variante e excluir do seu texto as que sejam mais "portuguesas" ou mais "brasileiras", mas também pode optar por uma forma de expressão num português mais internacional.

Os alinhamentos do e-PACT foram obtidos por meio de uma ferramenta de alinhamento – CLUE-Aligner (ver secção 4) e seguindo um conjunto de diretrizes – CLUE Paraphrase Alignment Guidelines (ver secção 5). O corpo e-PACT, as diretrizes de alinhamento de paráfrases, a ferramenta CLUE-Aligner e os recursos com os pares parafrásticos foram todos desenvolvidos no âmbito do projeto eSPERTo.

4 CLUE-Aligner – Ferramenta de Alinhamento

O CLUE-Aligner (**C**ross-**L**anguage **U**nit **E**licitation Aligner) (Barreiro, Raposo, and Luís 2016) é uma ferramenta de alinhamento interativa[8] projetada para a anotação e extração de unidades lexicais multipalavra e outras unidades frásicas em pares de frases de textos paralelos monolingues ou bilingues, ou seja, de textos parafrásticos (comparáveis) ou de corpos paralelos de traduções. Baseada no

[8] O CLUE-Aligner está disponível em `https://esperto.l2f.inesc-id.pt/esperto/aligner/index.pl?`

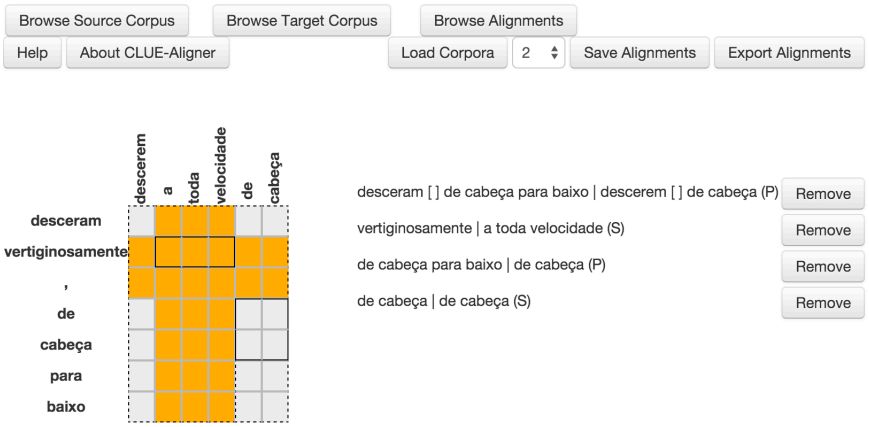

Figure 4: Interface do CLUE-Aligner com pares de unidades parafrásticas

alinhador Linear-B (Callison-Burch 2007; Callison-Burch and Bannard 2004), o CLUE-Aligner foi extendido de forma a permitir alinhar pares de unidades parafrásticas e de tradução quer sejam contíguas ou não-contíguas. A Figura 4 ilustra a versão atual da interface do CLUE-Aligner com o alinhamento de pares de unidades parafrásticas do corpo e-PACT contemplando unidades não-contíguas, representadas através de blocos também não-contíguos (*desceram de cabeça para baixo* = *descerem de cabeça*). Estas unidades surgem com elementos externos inseridos, ou seja, palavras que não fazem parte da unidade, os advérbios *vertiginosamente* e *a toda a velocidade*. A ferramenta permite o alinhamento de unidades individuais ou de unidades lexicais multipalavra mais pequenas (*de cabeça* (*para baixo*) = *de cabeça*) dentro das unidades lexicais multipalavra contíguas ou não contíguas mais largas. A secção 5 descreve com mais pormenor o processo de alinhamento de unidades parafrásticas e de tradução no CLUE-Aligner.

5 CLUE Alignment Guidelines – Diretrizes para o Alinhamento de Unidades Parafrásticas e de Tradução

As CLUE Alignment Guidelines são dois conjuntos de documentos com linhas diretrizes para o alinhamento de unidades lexicais multipalavra e outras unidades parafrásticas. Estas linhas diretrizes determinam cursos de ação relativamente ao alinhamento dessas unidades dependendo de elas serem encontradas e usadas em paráfrases – CLUE Paraphrase Alignment Guidelines – ou em tradução – CLUE Translation Alignment Guidelines. Relativamente a linhas diretrizes anteriores (Graça et al. 2008; Kruijff-Korbayová, Chvátalová, and Postolache 2006; Lam-

bert et al. 2005; Melamed 1998), as CLUE Alignment Guidelines tomam em linha de conta a anotação de alinhamentos de unidades não-contíguas, estando, por conseguinte, linguisticamente mais motivadas.

As CLUE Paraphrase Alignment Guidelines[9] resumem as recomendações e decisões mais importantes para o alinhamento de pares de unidades lexicais multipalavra e outras unidades em corpos monolingues/comparáveis, como por exemplo, o e-PACT, com o contraste entre as variantes PE e PB existente em frases paralelas dos livros de David Lodge extraídas do COMPARA (secção 5.1). As CLUE Translation Alignment Guidelines[10] resumem as recomendações e decisões mais importantes para o alinhamento de pares de unidades lexicais multipalavra e outras unidades em frases paralelas de corpos bilingues/multilingues, como por exemplo, o corpus de teste do Europarl (Koehn 2005) (secção 5.2). O objetivo das diretrizes para o alinhamento de unidades parafrásticas e de tradução é ajudar os linguistas a serem mais eficazes e consistentes na tarefa de anotação de diferentes tipos de fenómenos linguísticos representados nas diretrizes com exemplos reais e com a motivação inerente. Durante a tarefa de anotação, os linguistas corrigem manualmente alinhamentos automáticos errados e definem novos alinhamentos. Os pares lexicais/parafrásticos alinhados resultantes da tarefa constituem o conjunto de recursos que descrevemos a seguir.

5.1 Alinhamento de Unidades Parafrásticas

A tarefa de alinhamento que realizámos serviu para definir as linhas diretrizes de alinhamento de unidades parafrásticas – CLUE Paraphrase Alignment Guidelines – um trabalho em desenvolvimento que começou com alinhamentos entre o português europeu e o português do Brasil do corpo e-PACT, mas que pode e deve ser expandido para outros corpos e envolver outras variantes. Neste trabalho de investigação foram anotadas 268 frases alinhadas com unidades parafrásticas usando a ferramenta de alinhamento CLUE-Aligner, perfazendo um total de 6.233 pares, correspondendo a 4.016 pares únicos.

A Figura 5 ilustra o alinhamento das expressões verbais *estavam a divertir-se* e *estavam se divertindo*, que contém uma inserção adverbial, *apenas*. Este advérbio é um elemento externo à construção verbal e surge só de um dos lados, i.e., na frase em PB. Por outro lado, a expressão idiomática *por puro gozo* na frase em PE alinha com a expressão *pelo simples prazer da brincadeira* na frase em PB numa tradução

[9] As CLUE Paraphrase Alignment Guidelines estão disponíveis em `https://esperto.l2f.inesc-id.pt/clue-guidelines-paraphrases` .

[10] As CLUE Translation Alignment Guidelines estão disponíveis em `https://esperto.l2f.inesc-id.pt/clue-guidelines-translation` .

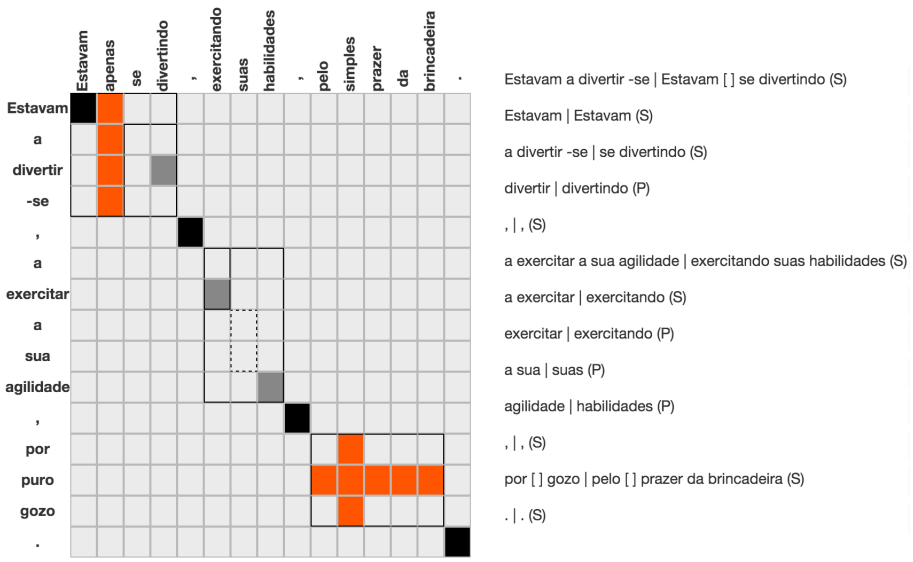

Figure 5: Alinhamento de unidades lexicais multipalavra não-contíguas.

mais livre.[11] Os pares de unidades parafrásticas resultantes estão listadas do lado direito da grelha de alinhamento com a informação se se trata de um alinhamento certo (S) ou de um alinhamento possível (P).[12] Os pares de unidades parafrásticas – Gold CLUE Paraphrases – estão disponíveis publicamente.[13]

5.2 Alinhamento de Unidades de Tradução Multilingue

Outra tarefa de alinhamento que realizámos visa obter pares de unidades de tradução multilingues, incluindo o alinhamento de várias categorias de unidades lexicais multipalavra contíguas e não-contíguas, unidades frásicas e de tradução, que constituem desafios importantes para a tradução automática de qualidade. A tarefa envolveu 400 frases abrangendo todas as combinações possíveis de textos paralelos em inglês, espanhol, francês e português do corpo de teste do Europarl, num total de 2.400 frases alinhadas. A coleção dourada dos alinhamentos anotados manualmente – Gold CLUE Translation – foi obtida com base nas linhas diretrizes

[11] O facto de cada uma das expressões terem sido encontradas nas traduções das diferentes variantes não significa que elas não sejam usadas pelos falantes das duas variantes.

[12] Um alinhamento S é um alinhamento ótimo e não-ambíguo, enquanto que um alinhamento P é um alinhamento não-ótimo e ambíguo. Uma definição de alinhamento certo e possível mais detalhada encontra-se em Barreiro (2016).

[13] `https://esperto.l2f.inesc-id.pt/gold-clue-paraphrases`

propostas – CLUE Translation Alignment Guidelines – para cada par de línguas e utilizando a nossa ferramenta de alinhamento – CLUE-Aligner.

A Figura 6 ilustra dois casos interessantes de alinhamentos bilingues inglês–português. Um dos pares, *os meus colegas – my colleagues*, são inserções no alinhamento dos verbos com preposição *Exorto* [ ] *a – I urge* [ ] *to*. Por outro lado, existe o alinhamento do verbo *acabarem* em português com a construção com verbo suporte *bring* [ ] *to a conclusion* em inglês, que contém a inserção do sintagma nominal *the interface specification*. Os pares de unidades de tradução resultantes estão listadas do lado direito da grelha de alinhamento com a informação se se trata de um alinhamento certo (S) ou de um alinhamento possível (P). Os pares de unidades de tradução – Gold CLUE Translation – estão disponíveis publicamente.[14]

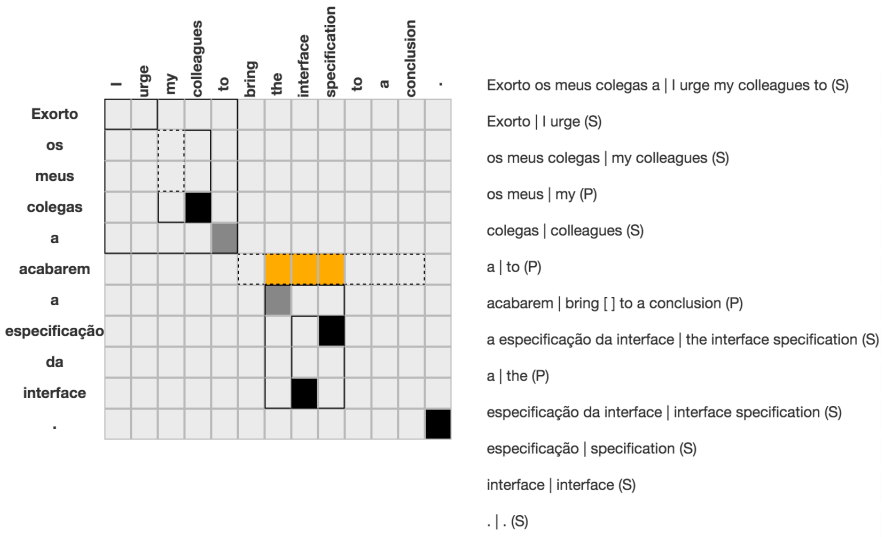

Figure 6: Alinhamentos de unidades não contíguas em pares de tradução

6 Conclusões e Trabalho Futuro

Neste artigo apresentámos dois recursos linguísticos e duas ferramentas que visam melhorar o nível de qualidade linguística da língua digital em aplicações de linguagem natural. A construção destes recursos representam esforços iniciais numa linha de investigação que reforça o papel do conhecimento linguístico no controlo

[14] https://esperto.l2f.inesc-id.pt/gold-clue-translation

de qualidade dos sistemas de processamento de linguagem natural e tecnologia da língua em geral. Para além de termos contribuído para o avanço da investigação nesta área tecnológica, reforçamos o apelo à necessidade real de investimento continuado para que se alcance um melhor desempenho dos sistemas de tecnologia da informação e comunicação, na mesma língua ou em línguas distintas. Destacamos o alargamento comunitário das coleções douradas (Gold CLUE Paraphrases e Gold CLUE Translation), o aperfeiçoamento e alargamento das linhas diretrizes de alinhamento (CLUE Alignment Guidelines) para a anotação de pares de unidades parafrásticas e de tradução e a construção de novos corpos que envolvam as diferentes variantes do português, não apenas para servirem como fonte de obtenção de paráfrases, mas também para que essas paráfrases "soem" naturais aos utilizadores do eSPERTo das várias comunidades de língua portuguesa. Assim sendo, propomos a exploração de corpos comparáveis de português dos PALOP, mas também de variedades regionais, variação linguística por diferentes faixas etárias, etc. de modo a oferecer ao utilizador uma maior variedade de escolhas, de acordo com as suas preferências ou objetivos da tarefa de parafraseamento. Depois da exploração do COMPARA, seria de utilidade explorar os corpos dos textos jornalísticos do CHAVE–PT e BR de 1994–1995 (não anotados), bem como as legendas de filmes e séries portuguesas do OpenSubtitles, entre outros. Um conjunto mais abrangente de textos ajudará a validar e explorar os nossos resultados iniciais e expandir os nossos recursos parafrásticos, não apenas a partir de gramáticas transformacionais que envolvem processos morfológicos derivacionais e outros processos linguísticos, mas também através do desenvolvimento de um motor híbrido de aquisição de paráfrases que integre conhecimento adquirido através de métodos estatísticos. Com um ângulo de visão mais abrangente, a base de conhecimento parafrástico será aplicada ao cenário da tradução automática onde tem várias utilidades, incluindo servir de base para a avaliação de sistemas de tradução automática, mas também servir como um módulo plug-in para reconhecer e integrar paráfrases no fluxo de trabalho da tradução.

Agradecimentos

Este trabalho foi financiado pela Fundação para a Ciência e Tecnologia (FCT), através do projeto exploratório eSPERTo (Ref. EXPL/MHC-LIN/2260/2013) e da bolsa de pós-doutoramento com a referência SFRH/BPD/91446/2012. Agradeço ao Tiago Luís e ao Francisco Raposo o desenvolvimento do CLUE-Aligner sem a qual os alinhamentos e as diretrizes não teriam sido possíveis de realizar.

References

Baptista, Jorge, Nuno Mamede, and Ilia Markov (2014). "Integrating a lexicon-grammar of verbal idioms in a Portuguese NLP system". PARSEME General Meeting, Athens, March 10-11, 2014 (poster session).

Barreiro, Anabela (2009). "Make it Simple with Paraphrases: Automated Paraphrasing for Authoring Aids and Machine Translation". PhD thesis. Porto, Portugal: Universidade do Porto.

— (2011). "SPIDER: A System for Paraphrasing in Document Editing and Revision — Applicability in Machine Translation Pre-editing". In: *Computational Linguistics and Intelligent Text Processing*. Ed. by Alexander Gelbukh. Vol. 6609. Lecture Notes in Computer Science. Springer, pp. 365–376.

— (2016). "Alignments of Multiwords and Phrasal Translation Units Inspired by the Logos Core Grammar". In: *Proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC 2016)*. Ed. by Nicoletta Calzolari et al. Portoro, Slovenia, pp. –.

Barreiro, Anabela and Luís Miguel Cabral (2009). "ReEscreve: a translator-friendly multi-purpose paraphrasing software tool". In: *Proceedings of the Workshop Beyond Translation Memories: New Tools for Translators, The Twelfth Machine Translation Summit*. Ed. by Marie-Josée Goulet et al. Château Laurier, Ottawa, Ontario, Canada, pp. 1–8.

Barreiro, Anabela and Cristina Mota (2016). "e-PACT: eSPERTo Paraphrase Aligned Corpus of EN-EP/BP Translations". In: *Proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC 2016)*. Ed. by Nicoletta Calzolari et al. Portoro, Slovenia, pp. –.

Barreiro, Anabela, Francisco Raposo, and Tiago Luís (2016). "CLUE-Aligner: An Alignment Tool to Annotate Pairs of Paraphrastic and Translation Units". In: *Proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC 2016)*. Ed. by Nicoletta Calzolari et al. Portoro, Slovenia, pp. –.

Barreiro, Anabela et al. (2011). "OpenLogos Rule-Based Machine Translation: Philosophy, Model, Resources and Customization". In: *Machine Translation* 25.2, pp. 107–126.

Barreiro, Anabela et al. (2013). "When Multiwords Go Bad in Machine Translation". In: *Proceedings of the Workshop on Multi-word Units in Machine Translation and Translation Technology, Machine Translation Summit XIV*, pp. 26–33.

Barreiro, Anabela et al. (2014a). "Linguistic Evaluation of Support Verb Construction Translations by OpenLogos and Google Translate". In: *Proceedings of the*

9th edition of the Language Resources and Evaluation Conference, 26-31 May, Reykjavik, Iceland.

Barreiro, Anabela et al. (2014b). "OpenLogos Semantico-Syntactic Knowledge-Rich Bilingual Dictionaries". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Ed. by Nicoletta Calzolari (Conference Chair) et al. Reykjavik, Iceland.

Callison-Burch, Chris (2007). "Paraphrasing and Translation". PhD thesis. Edinburgh, Scotland: University of Edinburgh.

Callison-Burch, Chris and Colin Bannard (2004). "Improving statistical translation through editing. European Association for Machine Translation (EAMT-04) Workshop". In: *European Association for Machine Translation*.

Frankenberg-Garcia, Ana and Diana Santos (2003). "Introducing COMPARA: the Portuguese-English Parallel Corpus". In: *Corpora in Translator Education*. Ed. by Federico Zanettin, Silvia Bernardini, and Dominic Stewart. Manchester: St. Jerome, pp. 71–87.

Graça, João et al. (2008). "Building a Golden Collection of Parallel Multi-Language Word Alignment". In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Ed. by Nicoletta Calzolari et al. ELRA.

Koehn, Philipp (2005). "EuroParl: A Parallel Corpus for Statistical Machine Translation". In: *Conference Proceedings: the tenth Machine Translation Summit*. AAMT, pp. 79–86.

Kruijff-Korbayová, Ivana, Klára Chvátalová, and Oana Postolache (2006). "Annotation Guidelines for Czech-English Word Alignment". In: *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pp. 1256–1261.

Lambert, Patrik et al. (2005). "Guidelines for Word Alignment Evaluation and Manual Alignment". In: *Language Resources and Evaluation* 39.4, pp. 267–285.

Melamed, I. Dan (1998). *Annotation Style Guide for the Blinker Project*. Tech. rep. IRCS.

Mota, Cristina, Paula Carvalho, and Anabela Barreiro (2016). "Port4NooJ v3.0: Integrated Linguistic Resources for Portuguese NLP". In: *Proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC 2016)*. Ed. by Nicoletta Calzolari et al. Portoro, Slovenia, pp. –.

Santos, Diana (2014). *Como estudar variantes do português e, ao mesmo tempo, construir um português internacional?* Presentation at Contact, Variation and Change: corpora development and analysis of Iberoromance language varieties workhop. Stockholm.

Santos, Diana (2015). “Portuguese language identity in the world: adventures and misadventures of an international language”. In: *Language - Nation - Identity: The questione della lingua in an Italian and non-Italian context*. Ed. by Elizaveta Khachaturyan. Cambridge Scholars Publishing, pp. 31–54.

Scott, Bernard (Bud) (2003). “The Logos Model: An Historical Perspective”. In: *Machine Translation* 18.1, pp. 1–72. ISSN: 0922-6567.

Scott, Bernard E. (1989). “The Logos System”. In: *MT Summit II*.

Silberztein, Max. *La formalisation des langues: l’approche NooJ*. Collection science cognitive et management des connaissances. ISBN: 9781784050535.

— (2016). *La formalisation des langues: l’approche NooJ*. Formalizing Natural Languages: the NooJ Approach. Wiley Eds.