DEZOTTI, L. C.; FERREIRA, A. D. Confrontando divergências estruturais na elaboração de diretrizes para alinhamentos de corpora paralelos latim-português. *ReVEL*, v. 22, n. 42, 2024. [www.revel.inf.br].

Confrontando divergências estruturais na elaboração de diretrizes para alinhamentos de *corpora* paralelos latim-português

Lucas Consolin Dezotti[1]
Anise D'Orange Ferreira[2]

lucas.dezotti@academico.ufpb.br
anise.ferreira@unesp.br

RESUMO: Este artigo descreve o processo de produção do primeiro alinhamento padrão ouro de textos paralelos bilíngues para o par latim-português. O artigo tem como foco a elaboração de diretrizes específicas de alinhamento e seu impacto na obtenção de um índice de acordo entre anotadores otimizado. As diversas fases do trabalho são detalhadas, incluindo as estratégias adotadas para confrontar as diversas divergências estruturais entre as duas línguas – parte inerente do trabalho de mapear e vincular segmentos de texto correspondentes em sentido –, em particular o processamento por níveis de análise linguística (Vauquois 1976) e a classificação das divergências de Dorr (1993). O resultado é o primeiro padrão ouro de alinhamento de tradução para o par latim-português, acompanhado de um conjunto de diretrizes que podem servir de guia para a produção de novos alinhamentos envolvendo este par de línguas. O processo como um todo, por sua vez, pode servir de modelo para a produção de diretrizes para outros pares de línguas.
PALAVRAS-CHAVE: *corpora* paralelos; diretrizes de alinhamento; latim; português.

ABSTRACT: This paper describes the steps undertaken to produce the first gold standard for Latin-Portuguese translation alignments. The paper focusses on how the alignment guidelines were created and its impact on improving the Inter Annotator Agreement (IAA) score. The working steps for producing both the alignment and the guidelines are detailed, discussing the strategies adopted in dealing with translation divergences – an intrinsic issue of mapping and linking texts fragments assumed to bear the same meaning –, namely, a method taken from transfer-based machine translation systems (Vauquois 1976) and a classification of those divergences (Dorr 1993). The result is the first gold standard for Latin-Portuguese translation alignments and a set of guidelines, providing a reliable base for new Latin-Portuguese translation alignments. Moreover, the entire process can be taken as an example of how to create guidelines for other translation pairs.
KEYWORDS: parallel corpora; translation alignment guidelines; Latin; Portuguese.

[1] Doutorando do Programa de Pós-Graduação em Linguística e Língua Portuguesa da Universidade Estadual Paulista (UNESP/FCL-Ar) | Professor do Departamento de Letras Clássicas e Vernáculas da Universidade Federal da Paraíba (UFPB).
[2] Doutora em Letras Clássicas pela Universidade de São Paulo (USP) | Professora do Departamento de Linguística, Literatura e Línguas Clássicas da Universidade Estadual Paulista (UNESP/FCL-Ar).

Introdução

O assombro e perplexidade causados pelo lançamento ao público do primeiro sistema gerativo baseado em um modelo de linguagem de grande escala ou *large language models* (LLM) – o já famoso ChatGPT[3] – é uma das manifestações visíveis da revolução trazida pelos modelos computacionais baseados em cálculos estatísticos para o desenvolvimento de sistemas de aprendizagem automática no âmbito do processamento de linguagem natural (PLN). Diferentemente dos modelos baseados em regras, que tentam espelhar o processamento da linguagem pela mente humana utilizando como parâmetros um léxico e regras de combinação, os modelos construídos sobre métodos estatísticos são independentes de regras: seu "conhecimento" é fundamentado em cálculos de probabilidades efetuados sobre um conjunto (cada vez mais amplo) de exemplos de uso. Tais avanços se verificam não apenas em sistemas geradores de material linguístico (capazes, por exemplo, de responder questões, criar resumos ou traduzir textos), mas também em sistemas classificadores (capazes de realizar diversos tipos de análise linguística, e.g. lematização, análise morfossintática, desambiguação de significados, reconhecimento de discursos de ódio e desinformação, análise de autoria e detecção de plágio, entre outros).[4]

Malgrado o suposto "tradicionalismo" associado à área dos Estudos Clássicos, a comunidade dos chamados "classicistas digitais" está entre as mais ativas na aplicação dos métodos e técnicas de PLN e da Linguística Computacional em suas pesquisas (Blackwell; Crane 2009: 51). Em particular, destaca-se o esforço colaborativo em construir uma *cyber*-infraestrutura formada por coleções digitais de textos e serviços de processamento computacional de línguas históricas, como o grego clássico e o latim (Blackwell; Crane 2009: 65–7). Entre os diversos projetos que participam desse esforço está o Ugarit[5] (Yousef *et al.* 2022-b), que provê uma ferramenta gratuita de alinhamento manual multilíngue e que tem investido no

[3] https://chat.openai.com/
[4] Para detalhes sobre esses sistemas, incluindo indicações de projetos colaborativos voltados para cada uma das tarefas mencionadas, ver Ježek e Sprugnoli (2022: 167–192).
[5] https://ugarit.ialigner.com/

desenvolvimento de uma *pipeline*[6] de alinhamento automático usando como dados de treinamento os alinhamentos manuais "padrão ouro" produzidos por especialistas (Yousef *et al.* 2022-a,c). Também há experiências de utilização desses alinhamentos tanto na produção de léxicos dinâmicos (Yousef; Berti 2015) quanto no âmbito do ensino de línguas (Palladino; Foradi; Yousef 2021).

No Brasil, o projeto pioneiro "Mutirão de Anotação de Corpus em Letras Clássicas Digitais", coordenado por Anise D'Orange Ferreira e executado de forma remota a partir da Faculdade de Ciências e Letras da UNESP, câmpus de Araraquara, ao longo de 2022, produziu anotações sintáticas e alinhamentos padrão ouro de textos em latim e grego clássico com suas respectivas traduções em português, visando seu uso como parâmetros de referência para o desenvolvimento de sistemas automatizados, a exemplo do Ugarit. Este artigo descreve o trabalho desenvolvido pela dupla de latinistas do projeto, que envolveu a elaboração de diretrizes específicas para o alinhamento de *corpora* paralelos latim-português, até então inexistentes, e que resultou no primeiro alinhamento padrão ouro para esse par de línguas. Assim, a próxima seção contextualiza o alinhamento de *corpora* paralelos no contexto geral das anotações, discutindo suas principais características e dificuldades. A Seção 2 descreve a primeira fase do trabalho de alinhamento no contexto do Projeto "Mutirão", feito sem diretrizes específicas, e faz uma avaliação quantitativa e qualitativa dos resultados obtidos. A seção 3 apresenta os princípios teóricos que nortearam a elaboração das diretrizes para alinhamentos latim-português, em particular uma abordagem comparativa segundo o nível de análise linguística inspirada no triângulo de Vauquois (1976) e a classificação das divergências de tradução de Dorr (1993). A Seção 4 descreve os resultados da segunda versão do alinhamento, já orientada pelas diretrizes. Por fim, apresentam-se algumas considerações finais sobre o processo, destacando suas vantagens, apontando suas limitações e prevendo futuros desdobramentos.

[6] Termo em inglês usado em computação para designar um sistema de processamento de dados organizado como uma sequência de módulos, em que cada módulo realiza uma tarefa específica e cujo resultado (*output*) serve de ponto de partida (*input*) para o módulo seguinte (Ježek; Sprugnoli 2023: 194).

1. Alinhamento e anotação de *corpus*

O termo **anotação** pode ser entendido de duas perspectivas, como produto e como processo: como produto, anotação é qualquer comentário associado a um determinado objeto (Wilcock 2009: 1); como processo, anotação é uma metodologia usada para adicionar informação a um documento na forma de metadados – isto é, dados a respeito de outros dados (Petrillo; Baycroft 2010: 2).

Uma anotação linguística, em particular, refere-se à adição de comentários sobre as propriedades linguísticas do texto anotado (Wilcock 2009: 1); nesse caso, os metadados registram informações que, em certa medida, coincidem com os níveis de análise linguística e se baseiam em técnicas de segmentação e identificação. De fato, partindo de uma concepção de texto como "sequência de elementos textuais que juntos formam uma cadeia (*string*)" (Tiedemann 2011: 7), os textos são divididos em segmentos de dimensões variadas, dependendo da finalidade da anotação. O tipo de segmentação mais comum opera no nível da palavra ou, mais tecnicamente, dos *tokens* – isto é, unidades mínimas normalmente delimitadas por espaços em branco (sinais de pontuação inclusos); a análise e o processamento permitem identificar, por exemplo, classes e morfologia das palavras (*PoS tagging*) e relações sintáticas entre elas (*parsing*), relações semânticas entre predicado e argumentos (*semantic analysis*), nomes de pessoas e lugares (*named entity recognition* ou NER), entre outras possibilidades (Wilcock 2009: 19-20).

Também é possível fazer uma anotação do tipo relacional, em que se mapeiam e identificam as "relações entre dois segmentos (ou duas anotações de segmentos)" (Ide et al. 2017: 85). Um exemplo é o alinhamento de textos paralelos ou bitextos, que são "documentos acompanhados de tradução para uma ou mais línguas diferentes" (Tiedemann, 2011: 1). Como as demais anotações, o alinhamento pode ser definido de duas perspectivas: como processo, refere-se ao trabalho de comparar dois textos em diferentes línguas visando encontrar as correspondências de tradução entre suas respectivas unidades textuais (Kay; Röscheisen, 1993); como produto, é um objeto que estabelece vínculos entre palavras ou períodos correspondentes de um mesmo texto em línguas diferentes (Ide et al. 2017: 85; Li; Kim; Lee 2008). Vale notar que o termo alinhamento se refere à estrutura completa do mapeamento, e não às ligações

individuais entre itens (Tiedemann 2011: 7): os pares individuais de segmentos alinhados são chamados segmentos de bitexto ou bissegmentos.

Considerando os recursos atualmente disponíveis de segmentação automática, os tipos mais utilizados são o alinhamento de períodos (baseado em *sentence boundary detection*) e o alinhamento de palavras (baseado em *tokenization*). A escolha depende de fatores como a natureza do texto fonte, as características da tradução e, especialmente, o uso que será feito dos dados produzidos, na medida em que cada aplicação apresenta demandas específicas com relação à forma dos dados anotados (Lambert et al. 2005: 268). Uma finalidade comum de projetos de anotação é produzir dados para o treinamento e avaliação de modelos computacionais (Ježek; Sprugnoli 2023: 134). *Corpora* paralelos alinhados, em particular, servem ao desenvolvimento de sistemas de aprendizagem automática supervisionada voltados para dois tipos de tarefas: a tradução automática baseada em modelos estatísticos (SMT) e a extração de léxicos bilíngues dinâmicos. No primeiro caso, alinhamentos baseados em sintagmas (no sentido de sequências de palavras, não necessariamente coincidentes com uma definição linguística do termo) têm produzido melhores resultados (Brown et al. 1993). No segundo caso, são preferíveis pares alinhados de palavras, segundo critérios lexicais, dada a necessidade de correspondências de alta precisão (Brew 1996; Och; Ney 2004: 418).

Sabe-se que o alinhamento de palavras oferece maiores dificuldades, principalmente em virtude de diferenças constitutivas entre os segmentos, que podem ser tanto ocasionais (causadas pelo estilo do tradutor, como no caso de traduções não literais, omissões ou inclusões) quanto sistemáticas, motivadas pelas divergências estruturais entre as línguas (Lambert et al. 2005: 276); a presença de palavras funcionais (e.g. artigos, preposições, partículas, verbos auxiliares) em apenas um dos lados do par, por exemplo, constitui um caso típico e sempre desafiador (Véronis; Langlais 2000: 382). Na prática do alinhamento, tais diferenças se expressam no tipo de vínculo entre os segmentos, que podem ser classificados com base no número de elementos que constituem cada lado do par: palavra-por-palavra (1-1); palavra-por-locução (1-n); locução-por-palavra (n-1); e locução-por-locução (n-n). O exemplo abaixo ilustra essa tipologia com uma proposta de alinhamento da famosa abertura da primeira *Catilinária* de Cícero:

(1)	*quo usque*	*tandem*	*abutere*	*Catilina*	*patientiā*	*nostrā*
	n-n	1-1	1-1	1-1	1-n	1-1
	até quando	afinal	abusarás	Catilina	da paciência	nossa

'afinal, até quando, Catilina, abusarás da nossa paciência.'

Com efeito, ainda que a maioria dos pares seja do tipo (1-1), pelo menos dois casos demandam uma escolha por parte do anotador: o primeiro é a expressão multipalavra 'quo usque', que inicia a frase, aqui alinhada em conjunto com 'até quando' (vínculo n-n); o segundo é a palavra 'patientiā', alinhada com o sintagma preposicionado 'da paciência' (vínculo 1-n).

No exemplo acima, o anotador baseou sua escolha em critérios idiomáticos e sintático;[7] outros critérios certamente produziriam resultados distintos. Isso mostra que o resultado de uma anotação manual depende, em grande medida, de um trabalho interpretativo do anotador – de preferência um especialista em anotação com competência em linguística (Ježek; Sprugnoli 2023: 134). Porém, para que os dados produzidos possam ser usados no treinamento de sistemas de aprendizagem automática, é necessário que eles sejam considerados reproduzíveis de modo similar por anotadores distintos (Ježek; Sprugnoli 2023: 154). Nesse sentido, uma das formas de medir a qualidade de uma anotação é calcular o nível de acordo entre os anotadores (*Inter Annotator Agreement*, IAA) a partir de uma mesma amostra de texto anotada de modo independente por cada um. Existem diversas métricas que realizam esse cálculo, todas tendo como resultado um valor numérico entre -1 e 1; no âmbito da Linguística Computacional, tradicionalmente o valor crítico mínimo é 0,80 (Ježek; Sprugnoli 2023: 157).

Para alcançar valores superiores a esse mínimo, é importante que os projetos de anotação tenham diretrizes (*annotation guidelines*) claras e objetivas para orientar os anotadores no trabalho de identificar os elementos e criar as associações apropriadas (Pustejovsky; Stubbs 2012: 24), informando sobre que dados devem ser

[7] No primeiro caso, a classificação de 'quo usque' como expressão multipalavra se justifica pelo critério da *idiomaticidade sintática* (Ramisch; Villavicencio 2022: 652), uma vez que a construção predominante do advérbio *usque* é precedendo sintagmas preposicionados e advérbios de lugar, e não sucedendo-os, como no exemplo acima; vale notar que a expressão é frequentemente grafada como palavra única, sem o espaço, o que pode ser visto como um índice de seu alto grau de gramaticalização. No segundo caso, o substantivo no caso ablativo *patientiā* funciona como argumento da forma verbal *abutere*, cujo equivalente em português ('abusarás') exige o complemento preposicionado.

anotados, quais associações devem ser feitas em determinadas circunstâncias e como lidar com casos atípicos (Petrillo; Baycroft 2010). Tais diretrizes costumam ser produzidas de forma dinâmica, com base em ciclos iterativos de aplicação, avaliação e revisão: a partir de um esquema inicial, anotações são feitas sobre uma amostra de texto e têm sua qualidade medida pelo IAA; caso o índice de acordo não seja satisfatório, os anotadores revisam o trabalho, procurando resolver os casos discrepantes, na chamada reconciliação; com base nessas resoluções, as diretrizes são aprimoradas com melhores explicações ou novos exemplos e servem de referência para o início de um novo ciclo (Ježek; Sprugnoli 2023: 135). Uma vez estabelecida a versão final das diretrizes, ela pode ser aplicada em larga escala para a obtenção do "padrão ouro" (*gold standard*), que é como se denomina o conjunto de dados anotados manualmente próprio para ser utilizado como base de treinamento de sistemas automatizados (Veronis; Langlais 2000: 387; Pustejovsky; Stubbs 2012: 24). O objetivo do padrão ouro é criar um patamar de referência elevado para mensurar o nível de performance do computador em relação às anotações manuais, isto é, o índice de acordo entre anotadores humanos torna-se um parâmetro-alvo para anotações automáticas geradas por computador (Petrillo; Baycroft 2010).

Foi justamente um processo como esse, de elaboração das diretrizes de alinhamento para bitextos latim-português e produção de um padrão ouro, que foi desenvolvido de forma pioneira pela dupla de latinistas brasileiros no âmbito do projeto "Mutirão de Anotação" da UNESP. Suas duas fases, correspondentes a dois ciclos de anotação, avaliação e revisão, serão descritas a seguir.

2. Primeira fase: alinhamento sem diretrizes específicas

A dupla de latinistas do projeto "Mutirão de anotação" da UNESP era formada por Lucas Consolin Dezotti e Jéssica Frutuoso Mello, ambos doutorandos de programas de pós-graduação em Linguística com experiência como professores de Língua e Literatura Latina no ensino superior.

Os textos a serem alinhados eram constituídos de um excerto[8] do *Resumo das Histórias Filípicas* escrito pelo historiador romano Justino (Seel 1972) e de sua respectiva tradução (então inédita), de autoria de um membro da dupla (Mello 2024).

A ferramenta de alinhamento utilizada foi o Ugarit, devido aos recursos de visualização e exportação dos dados que apresenta e à possibilidade de utilização do alinhamento produzido como base de treinamento de um sistema de alinhamento automático (cf. Introdução). A ferramenta é dotada de mecanismos de tokenização eficientes, que exigem um número mínimo de ações de pré-edição dos textos por parte dos anotadores.[9]

Dada a ausência de diretrizes específicas para o par latim-português,[10] nesta primeira fase os alinhamentos foram feitos segundo um princípio básico dos alinhamentos: "a correspondência entre dois segmentos deve envolver o menor número possível de palavras de cada texto, mas quantas forem necessárias para que os segmentos vinculados apresentem correspondência de sentido" (Lambert et al. 2005: 275). Do ponto de vista metodológico, os anotadores se conscientizaram das recomendações gerais para que o trabalho de anotação tivesse resultados consistentes (Petrillo; Baycroft 2010: 5-6), quais sejam:

(i) toda anotação deve ser fruto de trabalho individual; os anotadores não devem colaborar durante o processo (com exceção dos momentos específicos para discussão de discrepâncias e resolução das dúvidas);

(ii) toda anotação consiste em classificar/relacionar os fenômenos linguísticos definidos pelas diretrizes (não se trata, portanto, de anotar necessariamente todas as palavras); e

(iii) toda anotação implica em leitura, releitura e revisão das marcações e registro de comentários sobre o processo (como decisões consideradas

[8] Trata-se dos capítulos 4 a 6 do livro 18, composto por 45 *sentences*, 575 *types* e 797 *tokens* (excluindo pontuação). Os dados foram obtidos com o auxílio da ferramenta Voyant Tools: https://voyant-tools.org/?corpus=1d5bc922ebec865640ebd0da5a31dd11.

[9] No projeto "Mutirão", em particular, recomendou-se apenas a separação manual entre os sinais de aspas e as palavras a que se juntam graficamente (e.g. *"cur"* > *" cur "*). Outra possível necessidade, dependendo do estilo de alinhamento proposto, é a separação das partículas enclíticas tanto no texto-fonte (e.g. *longumque* > *longum que*) quanto no texto alvo (*retirá-lo* > *retirá -lo*), casos que preferimos resolver com vínculos do tipo n-n.

[10] Até o momento, o projeto Ugarit desenvolveu diretrizes específicas para três pares de línguas, todas tendo, como língua-fonte, o grego clássico e, como língua-alvo, respectivamente, o inglês (Palladino; Shamsian; Yousef 2022); o português (Ferreira; Reis 2022); e o latim (Palladino; Wright; Yousef 2022).

difíceis, críticas sobre as diretrizes, problemas com a ferramenta de anotação).

Uma vez produzida a primeira versão do alinhamento pelos dois anotadores[11], o IAA foi calculado[12] usando a fórmula do kappa de Cohen – uma métrica de consenso para medir o acordo entre observadores de dados categoriais (Landis; Koch 1977) – com base em uma matriz com as quantidades de pares estabelecidos por cada anotador conforme o tipo de vínculo (Tabela 1). O resultado foi um IAA de 0,72, considerado significativo, porém abaixo do valor mínimo aceitável de 0,80.

Justino 18.4–6 (v.1)		ANOTADOR 1				
		1-1	**1-n**	**n-1**	**n-n**	***total***
ANOT.2	**1-1**	385	8	0	6	*399*
	1-n	91	252	0	0	*343*
	n-1	0	0	6	0	*6*
	n-n	1	1	0	20	*22*
	total	*477*	*261*	*6*	*26*	*770*

Tabela 1. Matriz dos pares estabelecidos por cada anotador na primeira versão do alinhamento, categorizados por tipo de vínculo.

A discussão dos resultados evidenciou que a maior parte dos desacordos entre os anotadores estava relacionada com a decisão de incluir ou não palavras funcionais nos alinhamentos, que se reflete na alta quantidade (91) de pares anotados como 1-1 pelo anotador 1 (porque não as incluiu) e como 1-n pelo anotador 2 (porque as incluiu). Ficou evidente a necessidade de um tratamento padronizado da questão, que implicava em duas decisões: (1) se as palavras funcionais seriam incluídas ou não; e (2) caso fossem, com qual palavra de conteúdo elas seriam associadas.

Como vimos anteriormente, a primeira questão envolve considerar a finalidade do alinhamento: para sistemas de tradução automática, vínculos envolvendo sintagmas são melhores porque auxiliam os sistemas a produzir frases

[11] Os arquivos XML produzidos pela ferramenta de alinhamento podem ser consultados, respectivamente, em: https://github.com/lucascdz/psm/blob/main/stoa0167.stoa001.unesp-lat1-por1--Annotator1-version1.xml e https://github.com/lucascdz/psm/blob/main/stoa0167.stoa001.unesp-lat1-por1--Annotator2-version1.xml.

[12] O cálculo foi feito por meio de um *script* produzido em linguagem R que (1) processa os arquivos XML para extrair os pares de segmentos vinculados; (2) compara os tipos de vínculos utilizados pelos anotadores para cada par; e (3) gera a matriz de dados quantitativos e calcula o índice *kappa*. Cf. https://github.com/lucascdz/psm/blob/main/CalculateKappaFromXMLAlignmentFiles.R.

mais adequadas aos contextos; para a lexicografia, vínculos 1-1 são preferíveis porque capturam mais precisamente as equivalências. A decisão pela inclusão das palavras funcionais se deu por uma razão prática, qual seja, a maior possibilidade de reuso de um alinhamento de estilo "inclusivo". De fato, caso se necessite de vínculos 1-1, um alinhamento inclusivo pode ser transformado por meio de técnicas de pós-edição – uma vez que as palavras funcionais, sendo uma classe fechada, podem ser excluídas com base em uma simples lista de *stopwords*; já para um alinhamento formado exclusivamente por pares 1-1 não é possível restituir as palavras omitidas sem que o trabalho tenha de ser completamente refeito.

A segunda questão é de natureza analítica e se materializa em exemplos como *compellit deicere* ('obriga a jogar') ou *praebuit se comitem* ('oferece-se como companheiro), que exigem uma decisão quanto a vincular as palavras funcionais 'a' e 'como', respectivamente, em conjunto com os verbos ('obriga' e 'oferece') ou com os argumentos ('jogar' e 'companheiro'). Nesse caso, a resposta buscou inspiração em métodos desenvolvidos no contexto da tradução automática baseada em regras: o estudo das equivalências por níveis de análise linguística (Vauquois 1976) e a classificação das divergências de tradução (Dorr 1993). A próxima seção descreve esses conceitos e ilustra sua aplicação.

3. Diretrizes para um alinhamento consistente

As diretrizes de alinhamento de *corpora* paralelos latim-português aqui propostas visam orientar o anotador no estabelecimento consistente de vínculos entre segmentos de cada lado do par, particularmente quando a correspondência de sentido requer vínculos de segmentos assimétricos (1-n, n-1, n-n), auxiliando o anotador a empregar soluções similares para problemas similares.

Para tanto, além do princípio básico da correspondência de sentido e das recomendações metodológicas gerais, vistas acima, sugere-se aos anotadores adotar uma estratégia iterativa e hierárquica de refinamento dos vínculos, com base na constatação de que o mapeamento de correspondências entre segmentos maiores, como documentos, parágrafos e períodos, costuma ser mais óbvio do que entre segmentos menores, como sintagmas e palavras (Tiedemann 2011: 14-16). Na prática, trata-se de equiparar segmentos seguindo uma progressão que parte da estrutura

profunda em direção à superfície, visando encontrar o nível mais superficial em que a correspondência se verifica.

A ideia de mapear segmentos conforme o nível de análise linguística é inspirada na segunda geração de modelos de tradução automática desenvolvidos nas décadas de 1960 e 1970, baseados em um procedimento desenvolvido por Victor H. Yngve envolvendo três operações: (1) um período do texto na língua-fonte é analisado em termos de sua estrutura sintática; (2) essa estrutura é transferida para a língua-alvo com base em um mapeamento de equivalências estruturais; (3) um período na língua-alvo é produzido com base nessa estrutura (Vauquois 1976: 130–1). Esse modelo foi, posteriormente, extrapolado para incluir outros níveis de estratificação (nível morfêmico, estrutura sintática de superfície, estrutura sintática profunda, nível semântico-conceitual), em cada um dos quais a transferência é testada. Em outras palavras, os algoritmos fazem uma análise progressiva ascendente da língua-fonte, em termos de sua estrutura lexical, sintática e semântica, até uma representação idealizada; em seguida, a geração textual percorre o caminho inverso, numa marcha progressiva descendente, até encontrar as palavras na língua-alvo (cf. Ramisch 2012: 137). O chamado "triângulo de Vauquois" ilustra esse modelo (Figura 1).

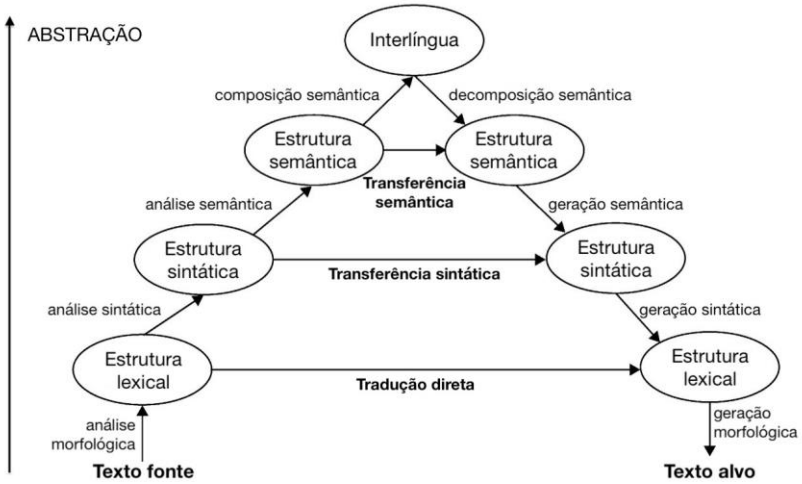

Figura 1. Triângulo de Vauquois. Adaptado de Door; Hovy; Levin (2006: 384).

Essa estratégia geral é secundada por diretrizes específicas que procuram fornecer orientação para um tratamento padronizado dos vários tipos de diferenças

normalmente encontradas[13] entre textos-fonte em latim e suas respectivas traduções para o português, que dificultam um alinhamento termo-a-termo.

Uma primeira distinção divide as diferenças entre texto-fonte e texto-alvo em dois grupos (Dorr 1993: p. 234): (1) o das **divergências** de tradução *(translation divergences),* quando a mesma informação é veiculada nos dois textos, porém através de estruturas distintas; e (2) o das **incompatibilidades** de tradução (*translation mismatches*), em que a própria informação veiculada é diferente nos dois textos. Ambos os casos serão discutidos e exemplificados no decorrer desta seção.

3.1. Divergências de tradução

Segundo Dorr (1993: 19), tem-se uma divergência de tradução quando "a tradução natural de uma língua em outra resulta numa forma muito diferente do original". Em outras palavras, tem-se uma divergência quando existe uma exceção no mapeamento entre categorias sintáticas e tipos léxico-semânticos entre as línguas com relação a uma representação formal, seja porque apresentam estruturas sintáticas distintas ou porque apresentam estruturas sintáticas semelhantes, porém com categorias básicas distintas (Dorr 1993: 240, 242).

A autora subdivide as divergências em duas classes. De um lado, divergências **morfossintáticas** são motivadas por propriedades sistemáticas associadas a cada língua (portanto, independentes dos itens lexicais presentes no enunciado) e consistem em diferenças relativas a ordem das palavras, formas de concordâncias, elipses e sujeito nulo, pronomes oblíquos e clíticos, etc. De outro lado, divergências **léxico-semânticas** são caracterizadas por propriedades determinadas pelo léxico e podem ser classificadas em sete tipos (Dorr 1993: 20–21, 240–7):

- Divergência **conflacional**: quando se verifica, em uma das línguas, a incorporação em um item lexical de componentes necessários para o significado (ou argumentos) de certa ação; na correspondência entre "dar um soco' e 'to punch', por exemplo, pode-se dizer que, em

[13] Alguns autores afirmam que a existência de diferenças entre textos na língua-fonte e na língua-alvo são mais uma constante do que uma exceção, mesmo quando se trata de línguas de mesma tipologia (Dorr; Hovy; Levin 2006: 388).

português, o componente ‘soco’ se realiza na superfície, ao passo que em inglês ele está incluído *(conflated)* no verbo.

- Divergência **estrutural**: quando há diferenças no tipo de sintagma que realiza o argumento; na correspondência entre ‘ele entrou na casa’ e ‘he entered the house’, por exemplo, o locativo se expressa por um sintagma preposicional em português (‘na casa’) e por um sintagma nominal em inglês (‘the house’).
- Divergência **temática**: quando há diferenças de mapeamento entre função sintática e papel temático dos elementos que realizam os argumentos; na correspondência entre ‘eu gosto de você’ e ‘me gustas tu’, por exemplo, o experienciador se realiza como sujeito em português (‘eu’) e objeto em espanhol (‘me’).
- Divergência **categorial**: quando palavras de uma língua são traduzidas por palavras de uma classe gramatical diferente na outra língua; na correspondência entre estar com ciúmes’ e ‘to be jealous’, por exemplo, o significado do substantivo ‘ciúme’ em português é codificado pelo adjetivo ‘jealous’ em inglês.
- Divergência **promocional/democional**: dois tipos que ocorrem quando há uma inversão na relação de dominância estrutural entre duas palavras semanticamente equivalentes de uma língua para outra; exemplo do primeiro tipo é a tradução de ‘John will probably come’ por ‘é provável que John venha’, em que a expressão da probabilidade é “promovida” de uma posição mais baixa em inglês (o adjunto adverbial ‘probably’) para uma posição mais alta em português (o predicado ‘é provável’); exemplo do segundo é a tradução da locução verbal ‘to run in’ por ‘entrar correndo’, em que o núcleo do predicado (‘run’) em inglês é “rebaixado” para uma posição de adjunto (‘correndo’) em português.
- Divergência **lexical**: ocorre no contexto das outras divergências e relaciona-se com as propriedades de realização e composição dos itens lexicais; uma vez que as divergências descritas acima afetam essas propriedades, pode ser vista como uma espécie de efeito colateral delas, e.g. na correspondência entre as expressões ‘dar um fim’ e *to put an end*

> se manifesta uma diferença no verbo suporte utilizado, respectivamente, em português (‘dar’) e em inglês (*put*).

Com base nessa classificação, é possível organizar os diversos casos de divergências entre o latim e o português e aplicar soluções padronizadas para os vínculos entre os segmentos, como se vê a seguir.

3.1.1. Divergências morfossintáticas entre o latim e o português. A esta categoria pertencem as diferenças relacionadas às categorias gramaticais (caso, número, gênero; tempo, modo, aspecto, voz), as quais, via de regra, não devem ser vistas como obstáculos para a vinculação entre itens lexicais das duas línguas.

Também pertence a esta categoria um bom número de casos em que um segmento em português inclui palavras funcionais que traduzem o sentido expresso por morfemas no texto latino; esses casos são normalmente resolvidos com vínculos do tipo 1-n. São exemplos:

– formas verbais perifrásticas (e.g. ‘scripserat’ || ‘tinha escrito’);[14]
– formas verbais acompanhadas de pronomes pessoais (e.g. ‘dico’ || ‘eu digo’);
– formas compostas por prefixação (e.g. ‘circumvolare’ || ‘voar ao redor’);
– adjetivos traduzidos por oração relativa (e.g. ‘longum’ || ‘que se prolonga’);
– adjetivos comparativos e superlativos (e.g. ‘fortior’ || ‘mais corajoso’).
– substantivos traduzidos por sintagmas nominais com presença de artigo e/ou pronomes em português (e.g. ‘nox’ || ‘a noite’, ‘fama’ || ‘sua fama’);
– substantivos traduzidos por sintagmas preposicionados em português (e.g. ‘vitae’ || ‘da vida’, ‘tibi’ || ‘para você’);
– palavras acompanhadas de formas enclíticas (e.g. ‘mentibusque’ || ‘e nas mentes’);
– preposições latinas traduzidas por locuções prepositivas em português (e.g. ‘apud’ || ‘na casa de’).
– orações reduzidas de infinitivo ou particípio em latim traduzidas por orações desenvolvidas em português, incluindo as respectivas conjunções, e.g.

[14] Em caso de formas perifrásticas ocorrerem em ambas as línguas, pode-se optar por dois vínculos independentes do tipo 1-1 (e.g. ‘scripta | est’ || ‘foi | escrita’).

(2)

agi	*seuere*	*uolo*
1-n	1-n	1-1
que seja tratada	com seriedade	quero

'quero que [a questão] seja tratada com seriedade.'

(3)

risum	*cupientes*	*tenere*	*nequimus*
1-n	1-n	1-1	1-n
o riso	embora o desejemos	conter	não podemos

'não podemos conter o riso, embora o desejemos.'

– pronomes relativos desempenhando ao mesmo tempo as funções de demonstrativo e relator (fenômeno comum em orações relativas com função substantiva em latim) traduzidos em português por pronomes distintos, e.g.

(4)

habes	*quod*	*facias*
1-1	1-n	1-1
tens	o que	fazer

(5)

sunt	*qui*	*foedera*	*rumpunt*
1-1	1-n	1-1	1-1
existem	aqueles que	acordos	quebram

– orações relativas em latim traduzidas por orações reduzidas em português, em que se vinculam os verbos junto com as respectivas palavras funcionais, e.g.

(6)

quae miretur	*multa*	*habet*
n-n	1-n	1-n
para admirar	muitas coisas	ela tem

'ela tem muitas coisas para admirar.'

3.1.2. Divergências estruturais entre o latim e o português. Como vimos, este tipo de divergência é semelhante às divergências morfossintáticas; a diferença é que são dependentes da valência do item lexical escolhido na tradução. Em todo caso, o critério de alinhamento é vincular os segmentos que cumprem a mesma função sintática, incluindo eventuais palavras funcionais em português,

resultando normalmente em um vínculo do tipo 1-n. O caso típico é quando complementos verbais, expressos em latim por uma forma flexionada, são traduzidos por sintagmas preposicionados em português, e.g.

(7)	*ferre*	*domum*	;	*carere*	*honore*	;	*fugere*	*urbem*
	1-1	1-n		1-n	1-n		1-1	1-n
	levar	para casa	;	estar privado	de honra	;	fugir	da cidade

3.1.3. Divergências temáticas entre o latim e o português. Quando há reposicionamento, na tradução, de um ou mais argumentos de determinado núcleo sintático, o critério de alinhamento é vincular os segmentos conforme cumpram o mesmo papel semântico, independente da forma que assumem na estrutura de superfície. Considerando o par latim-português, isso pode ocorrer em pelo menos três situações, a saber:

(a) quando o verbo escolhido pelo tradutor exige tal reposicionamento, e.g.

(8)	*non*	*ita*	*dis*	*placuit*
	1-1	1-1	1-n	1-1
	não	assim	os deuses	quiseram

‘os deuses não quiseram assim.’ (lit. ‘assim não agradou aos deuses.’)

(9)	*placitum est*	*mihi*	*ut…*
	n-1	1-1	1-1
	achei	eu	que

‘eu achei que…’ (lit. ‘me agradou que…’)

(b) quando se traduzem orações na voz passiva por orações na voz ativa, e.g.

(10)	*iuberis*	*a me*	*abire*
	1-1	n-1	1-n
	ordeno	eu	que te retires

‘eu ordeno que te retires’ (lit. ‘és ordenado por mim a te retirares’)

(c) nas construções do chamado dativo de posse latino, e.g.

(11)

huic	*filia*	*est*
1-1	1-n	1-1
ele	uma filha	tem

'ele tem uma filha' (lit. 'uma filha existe para ele')

3.1.4. Divergência "promocional" entre o latim e o português. O caso mais comum de divergência deste tipo é a tradução do chamado ablativo absoluto latino por uma oração desenvolvida em português, em que o particípio que cumpria uma função de adjunto adnominal passa a constituir o núcleo do predicado de uma oração adverbial (que pode estar na voz ativa ou na voz passiva), e.g.

(12)

hostibus	*profligatis*	*Caesar*
1-n	1-n	1-1
os inimigos	depois que foram derrotados	César

'depois que os inimigos foram derrotados, César...'
(lit. 'derrotados os inimigos, César...')

(13)

hostibus	*profligatis*	*Caesar*
1-n	1-n	1-1
os inimigos	após ter derrotado	César

'após ter derrotado os inimigos, César...'
(lit. 'derrotados os inimigos, César...')

3.1.5. Divergências categoriais entre o latim e o português. Há casos em que uma indicação semântica, expressa por certa classe de palavras em latim, é transferida para outra classe de palavras na tradução, e.g.

(14)

non potest	*utrumque*	*fieri*
n-n	1-n	1-n
é impossível	as duas coisas	que aconteçam

'é impossível que as duas coisas aconteçam.'
(lit. 'não pode acontecer as duas coisas')

3.1.6. Divergências lexicais entre o latim e o português. Nesta categoria se incluem os casos de expressões idiomáticas em latim traduzidas por uma palavra ou expressão de sentido equivalente em português (e.g. 'dare verba'|n-1|'enganar', 'dare lacrimas'|n-1|'chorar', 'dare poenam'|n-n|'ser punido', 'quam ob rem'|n-n|'por isso').

3.2. Incompatibilidades de tradução

É difícil prever casos de discrepância de informação entre o texto-fonte e a tradução. Um exemplo não incomum é o acréscimo, na tradução, de informações que estão apenas implícitas no texto-fonte – seja quando o sujeito oculto de um verbo é preenchido ou quando um pronome anafórico tem seu antecedente explicitado por um substantivo. Em ambos os casos, nenhum vínculo deve ser estabelecido, e.g.

(15)

hunc	*a*	*nostris*	*rationibus*	*seiunctum fore*
ø	1-1	1-1	1-1	n-n
Célio	de	nossos	interesses	se apartará

'Célio se apartará de nossos interesses.'
(lit. 'este se apartará de nossos interesses')

Além disso, via de regra, quaisquer palavras consideradas mal traduzidas devem ser deixadas sem alinhar, a critério do anotador.

3.3. Outras diferenças

3.3.1. Repetições. Se a tradução omite palavras que se repetem no texto-fonte, o alinhamento é feito com apenas uma das ocorrências, a critério do anotador, e.g.

(16)

et	*Philus*	*et*	*Manilius*	*adesset*
ø	1-1	1-1	1-1	1-n
–	Filo	e	Manílio	estavam presentes

'Filo e Manílio estavam presentes.'

3.3.2. Omissões. Se o tradutor utiliza uma única preposição para introduzir dois ou mais elementos de um sintagma preposicionado, considera-se que ela forma grupo apenas com a primeira palavra da sequência, e.g.

(17)

quod	*merito*	*atque*	*iure*	*contigit*
1-1	1-n	1-1	1-1	1-n
isso	por merecimento	e	justiça	se deu

'isso se deu por merecimento e justiça.'

3.3.3. Palavras não traduzidas. Se uma palavra é transcrita ou mantida na língua original na tradução, ela é deixada sem alinhar.

3.3.4. Sinais de pontuação. Não devem ser alinhados, com exceção do ponto-final usado em formas abreviadas, normalmente traduzidas por extenso em português (e.g. 'Kal .'|n-1|'Calendas'; 'M .'|n-1|'Marco'). Note-se que o ponto-final conta como *token*, por isso o vínculo é classificado como n-1.

4. Segunda fase: alinhamento com base em diretrizes específicas

As diretrizes acima especificadas foram aplicadas pelos dois anotadores do projeto Mutirão em uma nova anotação do mesmo trecho de Justino. Uma vez

terminada essa segunda versão do alinhamento,[15] a matriz com as quantidades de vínculos estabelecidos por cada anotador (Tabela 2) foi construída e o IAA foi calculado. Desta vez, o acordo entre os anotadores subiu para 0,95, considerado "perfeito" (Ježek; Sprugnoli 2023: 157).

Justino 18.4–6 (v.1)		ANOTADOR 1				
		1-1	**1-n**	**n-1**	**n-n**	***total***
ANOT.2	**1-1**	388	13	0	3	*404*
	1-n	1	342	0	1	*344*
	n-1	0	0	6	0	*6*
	n-n	0	0	0	22	*22*
	total	*389*	*355*	*6*	*26*	*776*

Tabela 2. Matriz dos pares estabelecidos por cada anotador na segunda versão do alinhamento, categorizados por tipo de vínculo.

Ainda assim, a discussão dos resultados evidenciou pelo menos dois pontos que necessitavam de padronização. De um lado, a hesitação do anotador 2 de incluir vínculos com pronomes possessivos (e.g. 'fama/loquebatur'||'sua fama/circulava') explica parte dos 13 vínculos com discrepância 1-n/1-1 entre os anotadores 1 e 2, respectivamente; nesse caso, decidiu-se por estender a abordagem inclusiva também aos tais pronomes, uma vez que constituem um recurso bastante comum nas traduções latim-português e conferem idiomaticidade ao texto-alvo. De outro lado, a segmentação de locuções prepositivas na tradução pelo anotador 1 (e.g. 'pro/salute'||'em nome/da salvação' em vez de 'em nome da/salvação') gerou três casos de discrepância n-n/1-1 entre eles, resolvidas conforme recomendam as diretrizes (cf. seção 3.1.1). Após estes últimos ajustes, o acordo entre os anotadores foi considerado suficiente e o alinhamento estabelecido como um exemplo de padrão ouro[16] para o par latim-português.

[15] Os arquivos XML desta segunda versão podem ser consultados, respectivamente, em: https://github.com/lucascdz/psm/blob/main/stoa0167.stoa001.unesp-lat1-por1--Annotator1-version2.xml e https://github.com/lucascdz/psm/blob/main/stoa0167.stoa001.unesp-lat1-por1--Annotator2-version2.xml.

[16] Disponível para visualização em: https://ugarit.ialigner.com/text.php?id=34072. Para o arquivo XML, cf. https://github.com/lucascdz/psm/blob/main/stoa0167.stoa001.unesp-lat1-por1--GoldStandard.xml.

Considerações finais

Neste artigo, descrevemos o trabalho realizado no âmbito do projeto "Mutirão de Anotação de Corpus em Letras Clássicas Digitais", que consistiu em uma experiência pioneira de alinhamento de *corpus* paralelo latim-português, envolveu a elaboração de diretrizes específicas (até então inexistentes) e resultou no primeiro padrão ouro de alinhamento de traduções deste par de línguas.

Os resultados apresentados levam a crer, em primeiro lugar, que as diretrizes elaboradas para alinhamentos de textos paralelos latim-português são eficientes, uma vez que foram responsáveis por elevar o acordo entre anotadores (IAA) de 0,72 para 0,95. Esse aparente sucesso parece corroborar a dupla estratégia que fundamentou a elaboração das diretrizes: de um lado, uma estratégia geral, iterativa e hierárquica, atrelada aos níveis de análise linguística, que visa prover os anotadores de recursos para uma tomada de decisão de forma independente e fundamentada; de outro lado, uma descrição de casos específicos baseada numa classificação de divergências de tradução, igualmente atrelada aos níveis de análise linguística, que favorece o emprego consistente dos mesmos tipos de vínculo para casos semelhantes. Presume-se que tais características tenham o potencial de conferir às diretrizes uma maior polivalência, podendo ser aplicadas a diferentes tipos de tradução (mais "literais" ou mais "literárias") e a diferentes tipos de finalidade (tradução automática ou lexicografia), embora uma avaliação mais precisa desse potencial dependa da realização de testes específicos.

Além disso, o processo dinâmico de elaboração dessas diretrizes, aqui descrito, incluindo os princípios e estratégias adotados, pode servir de modelo para trabalhos semelhantes para outros pares de línguas, haja vista a carência de diretrizes específicas de alinhamento envolvendo a Língua Portuguesa.

Por fim, a produção de novos alinhamentos latim-português pode se configurar em uma excelente ferramenta para evidenciar e quantificar as divergências de tradução típicas entre essas línguas, quer de forma manual, quer automatizada, com vantagens tanto para a pesquisa linguística quanto para o trabalho tradutório. O que este artigo oferece é apenas uma primeira prospecção, mas cujos resultados positivos parecem indicar que se vai na direção certa.

Agradecimentos

Os autores agradecem a todos os participantes do projeto “Mutirão de Anotação de Corpus em Letras Clássicas Digitais”, em especial à professora Jéssica Frutuoso Mello, por fornecer a tradução do texto anotado e pela incansável dedicação às atividades do grupo de latinistas.

Também agradecem à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES/PDSE n. 88887.716744/2022-00), ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq/Pq n. 310893/2022-4) e à Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP/e-Science n. 2022/09490-0) pelo apoio financeiro.

Referências Bibliográficas

BLACKWELL, Christopher; CRANE, Gregory. Cyberinfrastructure, the Scaife Digital Library and Classics in a Digital age. *Digital Humanities Quarterly*, v. 3, n. 1, 2009.

BREW, Chris; MCKELVIE, David. Word-pair extraction for lexicography. *In*: *Proceedings of the 2nd international conference on new methods in language processing*, 1996.

BROWN, Peter F.; LAI, Jennifer C.; MERCER, Robert L. Aligning sentences in parallel corpora. *In*: *ACL '91: Proceedings of the 29th annual meeting of the Association for Computational Linguistics*, 1993.

DORR, Bonnie Jean. *Machine translation:* a view from the lexicon. Cambridge, Mass.: The MIT Press, 1993.

DORR, Bonnie J. Machine translation divergences: a formal description and proposed solution. *Computational Linguistics*, v. 20, n. 4, 1994.

DORR, Bonnie; HOVY, E.; LEVIN, L. Machine Translation: Interlingual Methods. *In*: BROWN, Keith (ed.). *Encyclopedia of Language and Linguistics*. Second Edition. Amsterdam: Elsevier, 2006.

FERREIRA, Anise O.; REIS, Michel F. *Critérios ou Convenções de Alinhamento do Grego às Traduções em Português* (1.0) [Data set]. Zenodo, 2022. Disponível em: https://doi.org/10.5281/zenodo.7981097. Acesso em 28 nov. 2023.

IDE, Nancy; CHIARCOS, Christian; STEDE, Manfred; CASSIDY, Steve. Designing annotation schemes: from model to representation. *In*: IDE, Nancy; PUSTEJOVSKY, James. *Handbook of Linguistic Annotation*. Dordrecht: Springer, 2017.

JEŽEK, Elisabetta; SPRUGNOLI, Rachele. *Linguistica computazionale*: introduzione all'analisi automatica dei testi. Bologna: il Mulino, 2023.

KAY, Martin; RÖSCHEISEN, Martin. Text-translation alignment. *Computational Linguistics*, v. 19, n. 1, 1993.

LANDIS, J. Richard; KOCH, Gary G. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, v. 33, n. 1, 1977.

LI, Jin-Ji; KIM, Dong-Il; LEE, Jong-Hyeok. Annotation Guidelines for Chinese-Korean Word Alignment. *In*: *Proceedings of the 6th Conference on Language Resources and Evaluation (LREC 2008)*. Marrakech, 2008.

LAMBERT, Patrik; DE GISPERT, Adrià; BANCHS, Rafael; MARIÑO, José B. Guidelines for word alignment evaluation and manual alignment. *Language Resources and Evaluation*, v. 39, 2005.

MELLO, Jéssica F. *"Farei de você um exemplo":* tradução integral, com notas e comentários, do *Epítome das Histórias Filípicas*, de Justino. Tese de Doutorado. Programa de Pós-Graduação em Letras/Estudos Literários da Universidade Federal de Juiz de Fora, 2024.

OCH, Franz J.; NEY, Hermann. The alignment template approach to statistical machine translation. *Computational Linguistics*, v.30, n.4, 2004.

PALLADINO, Chiara; FORADI, Maryam; YOUSEF, Tariq. Translation Alignment for Historical Language Learning: a Case Study. *Digital Humanities Quarterly*, v. 15, n. 3, 2021.

PALLADINO, Chiara; SHAMSIAN, Farnoosh; YOUSEF, Tariq. *Translation Alignment:* Ancient Greek to English. Annotation Style Guide and Gold Standard. (1.0) [Data set]. Zenodo, 2022. Disponível em: https://doi.org/10.5281/zenodo.7362097. Acesso em 28 nov. 2023.

PALLADINO, Chiara; WRIGHT, David J; YOUSEF, Tariq. *Translation Alignment:* Ancient Greek to Latin. Annotation Style Guide and Gold Standard (1.0) [Data set]. Zenodo, 2022. Disponível em: https://doi.org/10.5281/zenodo.7981085. Acesso em 28 nov. 2023.

PETRILLO, Matthew; BAYCROFT, Jessica. *Introduction to Manual Annotation*. Fairview Research, 2010. Disponível em: https://gate.ac.uk/teamware/man-ann-intro.pdf. Acesso em 28 nov. 2023.

PUSTEJOVSKY, James; STUBBS, Amber. *Natural Language Annotation for Machine Learning*. Beijing: O'Reilly, 2012.

RAMISCH, Carlos. *A generic and open framework for multiword expressions treatment: from acquisition to applications*. Tese de Doutorado. Programa de Pós-Graduação em Computação da Universidade Federal do Rio Grande do Sul. Porto Alegre (Brasil)/Grenoble (France), 2012.

RAMISCH, Carlos; VILLAVICENCIO, Aline. Computational Treatment of Multiword Expressions. *In*: MITKOV, Ruslan (ed.). *The Oxford Handbook of Computational Linguistics*. Second Edition. New York: Oxford University Press, 2022.

SEEL, Otto (ed.). *M. Iuniani Iustini Epitoma Historiarum Philippicarum Pompei Trogi*. Post F. Ruehl. Stuttgart: Teubner, 1972.

TIEDEMANN, Jörg. *Bitext Alignment*. San Rafael: Morgan & Claypool, 2011.

VAUQUOIS, Bernard. Automatic Translation: A Survey of Different Approaches. *Statistical Methods in Linguistics*, v. 12, 1976.

VÉRONIS, Jean; LANGLAIS, Philippe. Evaluation of parallel text alignment systems: The ARCADE project. *In*: VÉRONIS, J. (ed.). *Parallel text processing:* alignment and use of translation corpora. Dordrecht: Springer, 2000.

WILCOCK, Graham. *Introduction to Linguistic Annotation and Text Analytics*. San Rafael: Morgan & Claypool, 2009.

YOUSEF, Tariq; BERTI, Monica. The Digital Fragmenta Historicorum Graecorum and the Ancient Greek-Latin Dynamic Lexicon. *In*: MAMBRINI, Francesco; PASSAROTTI, Marco; SPORLEDER, C. (eds.). *Proceedings of the Workshop on Corpus-Based Research in the Humanities (CRH)*, 10 December 2015 Warsaw, Poland. Warszawa: Institute of Computer Science, Polish Academy of Science, 2015.

YOUSEF, Tariq; PALLADINO, Chiara; SHAMSIAN, Farnoosh; FERREIRA, Anise O.; REIS, Michel F. An automatic model and Gold Standard for translation alignment of Ancient Greek. *In*: *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*. Marseille, 2022 (a).

YOUSEF, Tariq; PALLADINO, Chiara; SHAMSIAN, Farnoosh; FORADI, Maryam. Translation Alignment with UGARIT. *Information* v. 13, n. 65, 2022 (b).

YOUSEF, Tariq; PALLADINO, Chiara; WRIGHT, David J.; BERTI, Monica. Automatic Translation Alignment for Ancient Greek and Latin. *In*: *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2022)*. Marseille, 2022 (c).