

Alinhamento e etiquetagem de corpora paralelos no CLUVI (Corpus Linguístico da Universidade de Vigo)

José Luis Aguirre Moreno*

Alberto Álvarez Lugrís*

Iago Bragado Trigo*

Luz Castro Pena#

Xavier Gómez Guinovart*

Santiago González Lopo*

Angel López López#

José Ramom Pichel Campos#

Elena Sacau Fontenla*

Lara Santos Suárez*

* Seminário de Linguística Informática - Universidade de Vigo
imaxin software

1. Introdução¹

O CLUVI (Corpus Linguístico da Universidade de Vigo) é um corpus textual aberto de registos especializados de língua galega contemporânea oral e escrita. No seu estado actual de desenvolvimento, os textos da secção escrita do CLUVI pertencem a quatro registos especializados (dos âmbitos jurídico-administrativo, jornalístico, informático e literário) e a três “combinações” linguísticas relativamente ao galego (monolíngue galego, tradução galego-espanhol e tradução inglês-galego), e possuem uma extensão total aproximada de 4 milhões de palavras. Os textos do CLUVI repartem-se em quatro subcorpora, cada um deles com cerca de 1 milhão de palavras: o corpus paralelo TECTRA (CLUVI-1) de textos literários inglês-galego, o corpus paralelo LEGA (CLUVI-2) de textos jurídico-administrativos galego-espanhol, o corpus monolíngue XIGA (CLUVI-3) de textos sobre informática em galego e o corpus monolíngue MEGA (CLUVI-4) de linguagem dos meios de comunicação social. Os objectivos de processamento do CLUVI incluem a sua etiquetagem morfosintáctica completa e o alinhamento das equivalências oracionais dos dois corpora paralelos, a extracção de informação léxica, terminológica e fraseológica dos corpora etiquetados e a transferência de resultados para aplicações de tradução automática, extracção de informação e resumo de documentos, recuperação de informação na Internet e correctores gramaticais para processamento de textos. O alargamento do CLUVI com textos paralelos português-galego está em fase de elaboração.

¹ Este trabalho foi financiado pela Junta da Galiza, dentro dos projectos “Desenvolvimento e aplicação de técnicas de análise linguístico-computacional de corpora orais e escritos para o processamento do CLUVI (Corpus Linguístico da Universidade de Vigo)” (ref. PGIDT01PXI30203PR) e “Estudo e aquisição de recursos básicos de linguística computacional do galego para a elaboração e melhoria de aplicações informáticas de tecnologia linguística” (ref. PGIDT01TICC06E), e ainda pelo Ministerio da Ciência e da Tecnologia espanhol (MCYT) e o Fundo Europeu de Desenvolvimento Regional (FEDER), dentro do projecto “Processamento linguístico-computacional do Corpus Linguístico da Universidade de Vigo (CLUVI)” (ref. BFF2002-01385). Mais informação em <http://webs.uvigo.es/sli>.

Nesta exposição, apresentaremos o etiquetário (*tagset*) morfossintáctico utilizado no SLI (Seminário de Linguística Informática da Universidade de Vigo) para a anotação linguística do CLUVI, mostrando em concreto as soluções adoptadas para exploração dos corpora paralelos TECTRA e LEGA. Na exposição inclui-se uma descrição do etiquetário morfossintáctico para a língua galega elaborado pelo SLI de acordo com as directrizes padrão europeias estabelecidas pelo EAGLES (*Expert Advisory Group on Language Engineering Standards*) (Leech e Wilson 1996; Monachini e Calzolari 1996, 1999), adaptadas por primeira vez ao galego; e das correspondências entre o etiquetário SLI do galego e o etiquetário intermédio proposto pelo EAGLES (Leech e Wilson 1996) como representação linguisticamente neutral do conjunto de pares atributo-valor que descrevem a informação linguística codificada nas etiquetas. Apresenta-se também a metodologia desenvolvida conjuntamente pelo SLI e Imaxin Software para a lematização e etiquetagem morfossintáctica da secção galega do CLUVI. Finalmente, apresenta-se a aplicação web desenhada pelo SLI para a consulta pública dos corpora paralelos do CLUVI.

2. Os corpora paralelos do CLUVI: TECTRA e LEGA

O corpus TECTRA (Álvarez Lugris 2001) contém catorze romances em língua inglesa com as suas correspondentes traduções para o galego, perfazendo um total de 1.127.044 palavras, 551.878 das quais correspondem às catorze traduções galegas e 575.166 aos catorze originais ingleses. Nos apêndices deste trabalho pode-se ver a relação completa de originais e traduções agrupadas neste corpus. Numa fase inicial, os textos do TECTRA foram adquiridos mediante digitalização das obras impresas, posteriormente foram revistos (com o intuito de corrigir os inevitáveis erros de reconhecimento óptico de caracteres) e por fim foram etiquetados estruturalmente em XML, com informação sobre a obra, autor/a, tradutor/a, capítulo, secção, página, parágrafo e frase.

O corpus paralelo LEGA de textos jurídico-administrativos galego-espanhol contém 1 milhão de palavras tiradas dos 251 exemplares correspondentes ao ano 2001 das edições em galego e em espanhol do DOGA (Diário Oficial da Galiza) publicado pela Secretaria Geral da Conselheria da Presidência e Administração Pública da Junta da Galiza. Os textos do LEGA foram adquiridos a partir da versão electrónica em PDF dos originais, após um protocolo de depuração do texto fonte visando a optimização dos resultados do alinhamento.

O processamento lingüístico computacional dos corpora paralelos TECTRA e LEGA, dentro do projecto CLUVI do SLI, apresenta três vertentes diferenciadas: etiquetagem, alinhamento e exploração. Na etiquetagem dos textos em galego empregámos o padrão XML e o etiquetário morfossintáctico elaborado pelo SLI de acordo com as directrizes do EAGLES e descrito no seguinte capítulo. O sistema probabilístico para a etiquetagem e desambiguação utilizado no CLUVI, desenvolvido simultaneamente pelo SLI e Imaxin Software, usa de um léxico computacional do galego que contém as especificações morfossintácticas definidas no etiquetário do SLI.

Com o fim de permitir desenvolvimentos e aplicações baseadas nas correspondências gramaticais entre os textos etiquetados do galego e os textos etiquetados do inglês e do espanhol, estabeleceu-se a correspondência entre o etiquetário SLI para o galego e o

etiquetário EAGLES mediante o etiquetário intermédio padrão proposto também pelo EAGLES. Tanto o alinhamento quanto a exploração do CLUVI para a extracção de informação léxica beneficiam da anotação morfossintáctica e da sua correspondência bilingue.

O alinhamento dos textos paralelos armazena-se em formato TMX, por ser o padrão para a codificação em XML de memórias de tradução e de corpora paralelos independentemente da aplicação utilizada (Melby 2000). A consulta pública dos corpora paralelos do CLUVI, através do *site* do SLI, permite examinar e explorar equivalências bilingues galego-ínglês e galego-espanhol em textos reais com finalidades académicas de investigação e docência, e também como ferramenta para a tradução². É preciso assinalar o facto de o galego não dispor ainda de um dicionário bilingue inglês-galego-ínglês apto para a tradução, sendo que isto faz do CLUVI uma ferramenta de consulta imprescindível neste contexto. A seguir centraremos a exposição deste trabalho na descrição do etiquetário SLI para o galego e do etiquetário intermédio.

3. Etiquetário morfossintáctico do SLI

3.1. Desenho do etiquetário

Para a criação do etiquetário do SLI, no que diz respeito à identificação dos fenómenos gramaticais relevantes em galego, baseámo-nos principalmente na descrição gramatical de Álvarez, Regueira e Monteagudo (1986) e Álvarez e Xove (2002). Igualmente, adoptámos as propostas do EAGLES relativamente às categorias gramaticais e aos traços morfossintácticos que é preciso diferenciar. Para isso, não nos limitámos a seguir as directrizes gerais do EAGLES; pelo contrário, aplicámos estritamente o esquema de atributos e valores recomendado por Leech e Wilson (1996), adequando-o ao galego de modo análogo a como já se tem realizado com outras línguas, como o italiano e o alemão (Teufel 1996).

Partindo da divisão em categorias principais, obrigatórias segundo Leech e Wilson (1996), determinámos os traços morfossintácticos aplicáveis ao galego, prescindindo daqueles que não o são e agregando atributos ou valores quando preciso. Desta forma, não incluímos no etiquetário o valor “neuro” do atributo “género” para os substantivos comuns, uma vez que ele não é aplicável ao galego, mas acrescentámos o valor “mais-que-perfeito” ao atributo “tempo” dos verbos no modo indicativo para recolher a forma verbal simples que em galego se expressa “eu cantara”.

Um outro aspecto fundamental do desenho do etiquetário do galego é o estabelecimento das correspondências com o etiquetário intermédio do EAGLES. O etiquetário intermédio é uma representação linguisticamente neutral que descreve os traços linguísticos (descritos em forma de pares atributo-valor) incluídos num etiquetário, de modo a que se possam fazer corresponder facilmente com as marcas de outro conjunto de etiquetas (Leech e Wilson 1996). O etiquetário intermédio permite trabalhar com etiquetas definidas segundo a terminologia gramatical própria da língua galega e convertê-las automaticamente aos traços definidos no padrão do EAGLES. Aplicado ao corpus paralelo TECTRA, o etiquetário intermédio permite estabelecer de forma

² Para os diversos problemas atinentes à divulgação de corpora através da web, vid. Santos (1999).

automática a inequívoca correspondência entre a informação gramatical dos textos em galego e a dos textos em inglês, o que possibilita a exploração destas correspondências em processos linguístico-computacionais posteriores à etiquetagem, como o alinhamento dos bitextos ou a extração automática de informação léxica bilingue contextual e fraseológica. De forma mais geral, a correspondência do etiquetário galego com o etiquetário intermédio permite reutilizar os textos etiquetados em aplicações adaptadas ao padrão EAGLES.

3.2. Apresentação do etiquetário

Para cada categoria mostram-se três quadros:

1. Quadro das relações hierárquicas entre subcategorias, em que também se especificam as restrições na aplicação de atributos e valores. Esta descrição está baseada nas recomendações do EAGLES registadas por Monachini e Calzolari (1999), se bem que tratamos de assinalar a hierarquia entre as subcategorias e as restrições de aparição de uns valores com outros de forma unificada, não empregando um único quadro para cada atributo, mas para toda a categoria. Em cada coluna representamos um atributo. Em cada fila representamos as combinações dos diferentes valores admitidos. A hierarquia entre subcategorias indica-se por meio da agrupação gráfica dos valores no quadro. Quando um valor não é aplicável, deixa-se a casa correspondente vazia. Os atributos ou valores específicos da língua galega, agregados ao etiquetário, aparecem destacados a negrito.

2. Relação dos atributos e valores conforme com a numeração de Leech e Wilson (1996), após aplicada ao galego. Incluem-se os traços que são opcionais no padrão EAGLES se forem adoptados para a nossa língua. Assinalam-se entre parênteses os atributos e valores que não se consideram aplicáveis para o galego e a negrito os específicos desta língua, acrescentados ao etiquetário.

3. Inventário completo de etiquetas para cada categoria. Inclui-se uma palavra como exemplo, a etiqueta empregada na nossa codificação, a descrição gramatical correspondente à etiqueta e, ainda, a codificação da etiqueta intermédia correspondente segundo a numeração de atributos e valores apontada na relação anterior. Aqui também, os dígitos dos valores acrescentados para o galego são indicados a negrito.

A seguir exemplificamos a relação de etiquetas do etiquetário SLI para corpus de língua galega, agrupadas por categorias conforme com a orde e o formato da descrição padrão das recomendações do EAGLES, através das categorias substantivo, verbo e adjetivo³.

³ Pode-se consultar a relação completa de etiquetas do etiquetário em Aguirre et al. (2002, 2003).

3.2.1. Categoria substantivo

Cat = substantivo				
Atributos				
	Tipo	Género	Número	Forma
Valores	comum	masc, fem, (neu)	sg, pl	plena, dimin.
	próprio	masc, fem	sg	plena, dimin.
		masc, fem	pl	plena

Quadro 1: Hierarquia de subcategorias nominais

- (i) Tipo: 1. Comum. 2. Próprio.
(ii) Género: 1. Masculino. 2. Feminino. (3. Neutro.)
(iii) Número: 1. Singular. 2. Plural.
(iv) (Caso: 1. Nominativo...)
(v) **Forma: 1. Plena. 2. Diminutivo.**

Figura 1: Atributos e valores nominais

Exemplo	Etiqueta	Descrição	Etiqueta intermédia
<i>can</i>	NCMS	substantivo comum masculino singular	N11101
<i>folla</i>	NCFS	substantivo comum feminino singular	N12101
<i>homes</i>	NCMP	substantivo comum masculino plural	N11201
<i>mulleres</i>	NCFP	substantivo comum feminino plural	N12201
<i>canciño</i>	NCDMS	substantivo comum diminutivo masculino singular	N11102
<i>follíña</i>	NCDFS	substantivo comum diminutivo feminino singular	N12102
<i>homiños</i>	NCDMP	substantivo comum diminutivo masculino plural	N11202
<i>mulleriñas</i>	NCDFP	substantivo comum diminutivo feminino plural	N12202
<i>Aldán</i>	NPMS	substantivo próprio masculino singular	N21101
<i>Antía</i>	NPFS	substantivo próprio feminino singular	N22101
<i>Ancares</i>	NPMP	substantivo próprio masculino plural	N21201
<i>Burgas</i>	NPFP	substantivo próprio feminino plural	N22201
<i>Pedriño</i>	NPDMS	substantivo próprio diminutivo masculino singular	N21102
<i>Carmiña</i>	NPDFS	substantivo próprio diminutivo feminino singular	N22102

Quadro 2: Inventário de etiquetas nominais

3.2.2. Categoria verbo

Cat = verbo							
Atributos							
	Fin	Modo/ forma	Tempo	Pess	Nm	Gn	
Valores	finito	indicativo	pres, imperf, fut. pres., pret, mais-que-perf.	1,2, 2t,3	sg, pl		
		conj	pres, imperf, fut	1,2, 2t,3	sg, pl		
		imper		2, 2t	sg, pl		
		fut. pret.		1,2, 2t,3	sg, pl		
	não-finito	infinitivo					
		inf pess			1,2, 2t,3,	sg, pl	
		particípio				sg, pl	m, f
		gerúndio					
		ger pess			1,2	pl	
		(supino)					

Quadro 3: Hierarquia de subcategorias verbais

<p>(i) Pessoa: 1. Primeira, 2. Segunda. 3. Terceira. 4. Tratamento.</p> <p>(ii) Género: 1. Masculino. 2. Feminino. (3. Neutro.)</p> <p>(iii) Número: 1. Singular. 2. Plural.</p> <p>(iv) Finitude: 1. Finito. 2. Não-finito.</p> <p>(v) Forma verbal / Modo: 1. Indicativo. 2. Conjuntivo. 3. Imperativo. 4. Futuro do pretérito. 5. Infinitivo. 6. Particípio. 7. Gerúndio. (8. Supino). 9. Infinitivo pessoal. A. Gerúndio pessoal⁴.</p> <p>(vi) Tempo: 1. Presente. 2. Imperfeito. 3. Futuro. 4. Pretérito. 5. Mais-que-perfeito.</p> <p>(vii) Voz: 1. Activa. 2. Passiva.</p> <p>(viii) (Status: 1. Principal. 2. Auxiliar.)</p>
--

Figura 2: Atributos e valores verbais

Para facilitar a consulta do Quadro 4, dividimos o inventário de etiquetas verbais em grupos, segundo o tempo e modo verbais:

⁴ Forma verbal pouco frequente em galego, mas viva nalgumas zonas para a primeira e segunda pessoas do plural (Álvarez e Xove 2002: 319).

Indicativo presente			
<i>collo</i>	VIPRS1	indicativo presente primeira singular	V10111110
<i>colles</i>	VIPRS2	indicativo presente segunda singular	V20111110
<i>colle</i>	VIPRS2C	indicativo presente segunda singular de tratamento	V40111110
<i>colle</i>	VIPRS3	indicativo presente terceira singular	V30111110
<i>collemos</i>	VIPRP1	indicativo presente primeira plural	V10211110
<i>colledes</i>	VIPRP2	indicativo presente segunda plural	V20211110
<i>collen</i>	VIPRP2C	indicativo presente segunda plural de tratamento	V40211110
<i>collen</i>	VIPRP3	indicativo presente terceira plural	V30211110

Indicativo pretérito imperfeito			
<i>collia</i>	VICPS1	indicativo pretérito imperfeito primeira singular	V10111210
<i>collias</i>	VICPS2	indicativo pretérito imperfeito segunda singular	V20111210
<i>collia</i>	VICPS2C	indicativo pretérito imperfeito segunda singular de tratamento	V40111210
<i>collia</i>	VICPS3	indicativo pretérito imperfeito terceira singular	V30111210
<i>colliamos</i>	VICPP1	indicativo pretérito imperfeito primeira plural	V10211210
<i>colliades</i>	VICPP2	indicativo pretérito imperfeito segunda plural	V20211210
<i>collian</i>	VICPP2C	indicativo pretérito imperfeito segunda plural de tratamento	V40211210
<i>collian</i>	VICPP3	indicativo pretérito imperfeito terceira plural	V30211210

Indicativo pretérito			
<i>collin</i>	VIPES1	indicativo pretérito primeira singular	V10111410
<i>colliches</i>	VIPES2	indicativo pretérito segunda singular	V20111410
<i>colleu</i>	VIPES2C	indicativo pretérito segunda singular de tratamento	V40111410
<i>colleu</i>	VIPES3	indicativo pretérito terceira singular	V30111410
<i>collemos</i>	VIPEP1	indicativo pretérito primeira plural	V10211410
<i>collestes</i>	VIPEP2	indicativo pretérito segunda plural	V20211410
<i>colleron</i>	VIPEP2C	indicativo pretérito segunda plural de tratamento	V40211410
<i>colleron</i>	VIPEP3	indicativo pretérito terceira plural	V30211410

Indicativo mais-que-perfeito			
<i>collera</i>	VIAPS1	indicativo mais-que-perfeito primeira singular	V10111510
<i>colleras</i>	VIAPS2	indicativo mais-que-perfeito segunda singular	V20111510
<i>collera</i>	VIAPS2C	indicativo mais-que-perfeito segunda singular de tratamento	V40111510
<i>collera</i>	VIAPS3	indicativo mais-que-perfeito terceira singular	V30111510
<i>colleramos</i>	VIAPP1	indicativo mais-que-perfeito primeira plural	V10211510
<i>collerades</i>	VIAPP2	indicativo mais-que-perfeito segunda plural	V20211510
<i>colleran</i>	VIAPP2C	indicativo mais-que-perfeito segunda plural de tratamento	V40211510
<i>colleran</i>	VIAPP3	indicativo mais-que-perfeito terceira plural	V30211510

Indicativo futuro do presente			
<i>collerei</i>	VIFUS1	indicativo futuro do presente primeira singular	V10111310
<i>collerás</i>	VIFUS2	indicativo futuro do presente segunda singular	V20111310
<i>collerá</i>	VIFUS2C	indicativo futuro do presente segunda singular de tratamento	V40111310
<i>collerá</i>	VIFUS3	indicativo futuro do presente terceira singular	V30111310
<i>colleremos</i>	VIFUP1	indicativo futuro do presente primeira plural	V10211310
<i>colleredes</i>	VIFUP2	indicativo futuro do presente segunda plural	V20211310
<i>collerán</i>	VIFUP2C	indicativo futuro do presente segunda plural de tratamento	V40211310
<i>collerán</i>	VIFUP3	indicativo futuro do presente terceira plural	V30211310

Indicativo futuro do pretérito			
<i>collería</i>	VIPPS1	indicativo futuro do pretérito primeira singular	V10114010
<i>collerías</i>	VIPPS2	indicativo futuro do pretérito segunda singular	V20114010
<i>collería</i>	VIPPS2C	indicativo futuro do pretérito segunda singular de tratamento	V40114010
<i>collería</i>	VIPPS3	indicativo futuro do pretérito terceira singular	V30114010
<i>colleríamos</i>	VIPPP1	indicativo futuro do pretérito primeira plural	V10214010
<i>colleríades</i>	VIPPP2	indicativo futuro do pretérito segunda plural	V20214010
<i>collerían</i>	VIPPP2C	indicativo futuro do pretérito segunda plural de tratamento	V40214010
<i>collerían</i>	VIPPP3	indicativo futuro do pretérito terceira plural	V30214010

Conjuntivo presente			
<i>colla</i>	VSPRS1	conjuntivo presente primeira singular	V10112110
<i>collas</i>	VSPRS2	conjuntivo presente segunda singular	V20112110
<i>colla</i>	VSPRS2C	conjuntivo presente segunda singular de tratamento	V40112110
<i>colla</i>	VSPRS3	conjuntivo presente terceira singular	V30112110
<i>collamos</i>	VSPRP1	conjuntivo presente primeira plural	V10212110
<i>collades</i>	VSPRP2	conjuntivo presente segunda plural	V20212110
<i>collan</i>	VSPRP2C	conjuntivo presente segunda plural de tratamento	V40212110
<i>collan</i>	VSPRP3	conjuntivo presente terceira plural	V30212110

Conjuntivo pretérito			
<i>collese</i>	VSPES1	conjuntivo pretérito primeira singular	V10112210
<i>colleses</i>	VSPES2	conjuntivo pretérito segunda singular	V20112210
<i>collese</i>	VSPES2C	conjuntivo pretérito segunda singular de tratamento	V40112210
<i>collese</i>	VSPES3	conjuntivo pretérito terceira singular	V30112210
<i>collesemos</i>	VSPEP1	conjuntivo pretérito primeira plural	V10212210
<i>colleledes</i>	VSPEP2	conjuntivo pretérito segunda plural	V20212210
<i>collesen</i>	VSPEP2C	conjuntivo pretérito segunda plural de tratamento	V40212210
<i>collesen</i>	VSPEP3	conjuntivo pretérito terceira plural	V30212210

Conjuntivo futuro			
<i>coller</i>	VSFUS1	conjuntivo futuro primeira singular	V10112310
<i>colleres</i>	VSFUS2	conjuntivo futuro segunda singular	V20112310
<i>coller</i>	VSFUS2C	conjuntivo futuro segunda singular de tratamento	V40112310
<i>coller</i>	VSFUS3	conjuntivo futuro terceira singular	V30112310
<i>collermos</i>	VSFUP1	conjuntivo futuro primeira plural	V10212310
<i>collerdes</i>	VSFUP2	conjuntivo futuro segunda plural	V20212310
<i>colleren</i>	VSFUP2C	conjuntivo futuro segunda plural de tratamento	V40212310
<i>colleren</i>	VSFUP3	conjuntivo futuro terceira plural	V30212310

Imperativo			
<i>colle</i>	VIMPS2	imperativo segunda singular	V20113010
<i>colla</i>	VIMPS2C	imperativo segunda singular de tratamento	V40113010
<i>collede</i>	VIMPP2	imperativo segunda plural	V20213010
<i>collan</i>	VIMPP2C	imperativo segunda plural de tratamento	V40213010

Formas não-finitas			
<i>coller</i>	VINFCS1	infinitivo pessoal primeira singular	V10129010
<i>colleres</i>	VINFCS2	infinitivo pessoal segunda singular	V20129010
<i>coller</i>	VINFCS2C	infinitivo pessoal segunda singular de tratamento	V40129010
<i>coller</i>	VINFCS3	infinitivo pessoal terceira singular	V30129010
<i>collermos</i>	VINFCP1	infinitivo pessoal primeira plural	V10229010
<i>collerdes</i>	VINFCP2	infinitivo pessoal segunda plural	V20229010
<i>colleren</i>	VINFCP2C	infinitivo pessoal segunda plural de tratamento	V40229010
<i>colleren</i>	VINFCP3	infinitivo pessoal terceira plural	V30229010
<i>coller</i>	VINF	Infinitivo	V00025010
<i>collendo</i>	VGER	Gerúndio	V00027010
<i>colléndomos</i>	VGERCP1	gerúndio pessoal primeira plural	V1022A010
<i>colléndodes</i>	VGERCP2	gerúndio pessoal segunda plural	V2022A010
<i>collido</i>	VPARMS	particípio masculino singular	V01126010
<i>collida</i>	VPARFS	particípio feminino singular	V02126010
<i>collidos</i>	VPARMP	particípio masculino plural	V01226010
<i>collidas</i>	VPARFP	particípio feminino plural	V02226010

Quadro 4: Inventário de etiquetas verbais

3.2.3. Categoria adjetivo

Cat = adjetivo				
	Atributos			
	Grau	Gén	Núm	<i>Forma</i>
Valores	normal	m, f	sg, pl	plena, diminutivo
	normal	m, f	sg	apocopado
	superlativo, comparativo	m, f	sg, pl	plena

Quadro 5: Hierarquia de subcategorias do adjetivo

- (i) Grau: 1. Normal. 2. Comparativo. 3. Superlativo.
(ii) Género: 1. Masculino. 2. Feminino. (3. Neutro.)
(iii) Número: 1. Singular. 2. Plural.
(iv) (Caso: 1. Nominativo ...)
(v) **Forma: 1. Plena. 2. Diminutivo. 3. Apocopado.**

Figura 3: Atributos e valores do adjetivo

<i>novo</i>	AXMS	adjectivo masculino singular	AJ11101
<i>nova</i>	AXFS	adjectivo feminino singular	AJ12101
<i>novos</i>	AXMP	adjectivo masculino plural	AJ11201
<i>novas</i>	AXFP	adjectivo feminino plural	AJ12201
<i>noviño</i>	AXDMS	adjectivo diminutivo masculino singular	AJ11102
<i>noviña</i>	AXDFS	adjectivo diminutivo feminino singular	AJ12102
<i>noviños</i>	AXDMP	adjectivo diminutivo masculino plural	AJ11202
<i>noviñas</i>	AXDFP	adjectivo diminutivo feminino plural	AJ12202
<i>novísimo</i>	AXSMS	adjectivo superlativo masculino singular	AJ31101

<i>novísima</i>	AXSFS	adjectivo superlativo feminino singular	AJ32101
<i>novísimos</i>	AXSMP	adjectivo superlativo masculino plural	AJ31201
<i>novísimas</i>	AXSFP	adjectivo superlativo feminino plural	AJ32201
<i>mellor</i>	AXCMS	adjectivo comparativo masculino singular	AJ21101
<i>mellor</i>	AXCFS	adjectivo comparativo feminino singular	AJ22101
<i>mellores</i>	AXCMP	adjectivo comparativo masculino plural	AJ21201
<i>mellores</i>	AXCFP	adjectivo comparativo feminino plural	AJ22201
<i>gran</i>	AXAPMS	adjectivo apocopado masculino singular	AJ11103
<i>gran</i>	AXAPFS	adjectivo apocopado feminino singular	AJ12103

Quadro 6: Inventário de etiquetas do adjectivo⁵

3.3. Etiquetas compostas

Há algumas características do galego que requerem um tratamento específico na anotação morfossintáctica de corpora nesta língua. Por um lado, a “fusão” de duas palavras numa só palavra ortográfica, própria das contracções, dos enclíticos e da segunda forma do artigo (“lo”, “la”, “los”, “las”; v. gr. “Terán que reconece-lo seu mérito”). No esquema de codificação do SLI, este conjunto de fenómenos recebem uma etiqueta “composta”, formada pela etiqueta da primeira palavra seguida da(s) etiqueta(s) correspondente(s) à(s) palavra(s) ligada(s), todas elas separadas pelo signo “_”. Portanto, a contracção da preposição “en” com o pronome pessoal masculino singular de 3ª pessoa “el” (isto é, “nel”) recebe a etiqueta “PREP_PPMS3”; a forma verbal com enclítico “dixome” recebe a etiqueta composta “VIPES3_PPS1A”, constituída pelas anotações para “verbo indicativo pretérito terceira singular” e “pronome pessoal átono singular primeira”; a “acercóuselle”, com dois enclíticos, corresponde a etiqueta “VIPES3_PPS3AR_PPS3AD”, composta pelas anotações para “verbo indicativo pretérito terceira singular”, “pronome pessoal singular átono terceira reflexivo” e “pronome pessoal átono singular terceira dativo”, ao passo que a “dixomo”, com dois enclíticos (“me” e “o”) em amálgama, atribui-se a etiqueta composta “VIPES3_PPS1A_PPMS3AA”, com as etiquetas correspondentes a “verbo indicativo pretérito terceira singular”, “pronome pessoal átono singular primeira” e “pronome pessoal átono masculino singular terceira acusativo”.

Quanto às segundas formas do artigo, utiliza-se o mesmo sistema para a composição das etiquetas compostas, diferenciando a segunda forma do artigo por meio da adição da anotação “-2” na sua etiqueta. Portanto, “bebe-lo” (“beber”+“o” artigo) codifica-se através da etiqueta composta VINF_ARDMS-2, formada pelas etiquetas de “verbo infinitivo” e de “artigo determinado masculino singular segunda forma”. As segundas formas do artigo também podem unir-se com hífen a um pronome enclítico, como em “gústalle-lo” (“gusta”+“lles”+“o”), que receberia a marca “VIPRS3_PPP3AD_ARDMS-2”, composta pelas etiquetas de “verbo indicativo presente terceira singular”, “pronome pessoal átono plural terceira dativo” e “artigo determinado masculino singular segunda forma”. Utiliza-se o mesmo sistema de anotação com as segundas e terceiras formas dos pronomes pessoais átonos acusativos de terceira pessoa, de modo que “bebelo” (“beber”+“o” pronome) codifica-se como VINF_PPMS3AA-2 (“verbo infinitivo”+“pronome pessoal átono masculino singular terceira acusativo segunda forma”), e “colleuna” (“colleu”+“a”) como

⁵ No grau comparativo e nas formas apocopadas dos adjectivos, optamos por diferenciar os valores de género embora essa diferenciação não se reflecta morfológicamente.

VIPES3_PPFS3AA-3 (“verbo indicativo pretérito terceira singular”+“pronome pessoal átono feminino singular terceira acusativo terceira forma”).

Além disto, outro traço do galego que exige um tratamento específico na anotação morfossintáctica são as locuções, nomeadamente as locuções prepositivas, conjuntivas e adverbiais. Na sua codificação opta-se por uma solução na linha do proposto por Sampson (1995), isto é, quando uma palavra faz parte de uma locução, atribui-se-lhe a etiqueta correspondente à sua categoria, seguida do signo “_”, da etiqueta correspondente à categoria da locução, e de um número de dois algarismos sendo o primeiro o número de palavras que integram a locução e o segundo, o número correspondente ao lugar que ocupa a palavra dentro da locução. Por exemplo, na locução prepositiva “cara a”, a “cara” atribui-se a etiqueta composta “NCFS_PREP21”, ao passo que a “a” lhe corresponde a etiqueta “PREP_PREP22”.

4. DTD do CLUVI-TMX

Os textos etiquetados no CLUVI seguem o padrão XML e incluem, no caso dos corpora paralelos, informação morfossintáctica e informação sobre equivalências de tradução. A DTD CLUVI-TMX da secção paralela do CLUVI é uma versão modificada do padrão TMX. Nesta versão modificada a informação morfossintáctica fica enquadrada por um elemento <ling> empregado para etiquetar todas as palavras e signos de pontuação dos elementos <seg> da estrutura TMX. Eis a definição do tipo de documento CLUVI-TMX:

```
<!-- DTD do CLUVI-TMX>
<!ELEMENT cluvi-tmx (header, body) >
<!ATTLIST cluvi-tmx
  version CDATA #REQUIRED >
<!ELEMENT header (#PCDATA)>
<!ATTLIST header
  creationtool CDATA #REQUIRED
  creationtoolversion CDATA #REQUIRED
  segtype (block|paragraph|sentence|phrase) #REQUIRED
  o-tmf CDATA #REQUIRED
  adminlang CDATA #REQUIRED
  srclang CDATA #REQUIRED
  datatype CDATA #REQUIRED >
<!ELEMENT body (tu*) >
<!ELEMENT tu (tuv+) >
<!ELEMENT tuv (seg) >
<!ATTLIST tuv
  lang CDATA #REQUIRED>
<!ELEMENT seg (ling+)>
<!ELEMENT ling (mor, ort)>
<!ELEMENT mor EMPTY>
<!ATTLIST mor
  cat (ARDFP|ARDFS|ARDMP|ARDMS...) #REQUIRED
  lema CDATA #REQUIRED>
  lema2 CDATA #IMPLIED
<!ELEMENT ort (#PCDATA)>
```

Desta forma, incorporamos a informação morfossintáctica como um alargamento da estrutura interna do elemento <seg> do formato TMX, como se pode observar na seguinte unidade de tradução tirada do TECTRA (no exemplo, elimina-se a informação morfossintáctica relativa à variante em língua inglesa da unidade de tradução para facilitar a sua compreensão):

```

<tu> <tuv lang="en"> <seg>In the town they tell the story of the great
pearl.</seg> </tuv> <tuv lang="gl"> <seg><ling> <mor lema="en" lema2="o"
cat="PREP_ARDFS"/> <ort>Na</ort> </ling> <ling> <mor lema="cidade"
cat="NCFS"/> <ort>cidade</ort> </ling> <ling> <mor lema="contar" lema2="se"
cat="VIPRS3_PPS3AR"/> <ort>cóntase</ort> </ling> <ling> <mor lema="o"
cat="ARDFS"/> <ort>a</ort> </ling> <ling> <mor lema="historia" cat="NCFS"/>
<ort>historia</ort> </ling> <ling> <mor lema="de" lema2="o"
cat="PREP_ARDFS"/> <ort>da</ort> </ling> <ling> <mor lema="grande"
cat="AXAPFS"/> <ort>gran</ort> </ling> <ling> <mor lema="perla" cat="NCFS"/>
<ort>perla</ort> </ling> <ling> <mor lema="." cat="PUNTO"/> <ort>.</ort>
</ling> </seg> </tuv> </tu>

```

5. Consulta do CLUVI na Internet

A aplicação desenhada pelo SLI para permitir a consulta pública dos corpora paralelos do CLUVI através da web⁶ segue as directrizes do corpus COMPARA⁷ relativamente às suas funcionalidades (Frankenberg-Garcia e Santos, no prelo). No entanto, os corpora paralelos TECTRA e LEGA do CLUVI acessíveis através do site do SLI estão em formato XML (na especificação CLUVI-TMX descrita no ponto 4 deste trabalho), e a ferramenta de consulta desenhada pelo SLI está criada especificamente para realizar pesquisas bilíngues em textos que respeitem as especificações TMX do XML (incluído o formato CLUVI-TMX).

O resultado das consultas possui uma aparência muito familiar para as pessoas que já alguma vez utilizaram o COMPARA, o que julgamos que pode facilitar a consulta profissional do CLUVI. Eis, a modo de exemplo, um excerto do resultado de uma consulta simples:

Palabra buscada: **cidade**

Equivalencia de traducción: **galego > inglés**

Total de equivalencias de traducción atopadas: **44**

1- PER (1)	Na cidade cóntase a historia da gran perla, de como foi atopada e de como foi perdida de novo.	"In the town they tell the story of the great pearl _ how it was found and how it was lost again.
2- PER (6)	En calquera caso, na cidade cóntase que...	In any case, they say in the town that..."
3- PER (127)	¿E el que había vir facer alí cando tiña traballo de máis entre a xente rica que vivía nas casas de pedra e de cemento da cidade ?	Why should he, when he had more than he could do to take care of the rich people who lived in the stone and plaster houses of the town.
4- PER (139)	Formaban unha apurada procesión de pasos silenciosos cara ó centro da cidade . Primeiro, Juana e Kino, detrás Juan Tomás e Apolonia soletreando coa súa gran barriga a ritmo marcial, e logo tódolos veciños cos nenos marchando a ámbolos lados.	They made a quick soft-footed procession into the center of the town, first Juana and Kino, and behind them Juan Tomás and Apolonia, her big stomach jiggling with the strenuous pace, then all the neighbors with the children trotting on the flanks.

⁶ Os corpora paralelos do CLUVI están dispoñibles em <http://sli.uvigo.es/CLUVI/>.

⁷ Dispoñível em <http://www.linguateca.pt/COMPARA/>.

6. Conclusões

Neste artigo apresentámos um etiquetário morfossintáctico normalizado (conforme aos padrões europeus definidos pelo EAGLES) para etiquetar corpora linguísticos de língua galega em aplicações monolíngues e plurilíngues. Apresentámos também a metodologia desenvolvida conjuntamente pelo SLI e Imaxin Software para a lematização e etiquetagem morfossintáctica da secção galega do CLUVI, e ainda a aplicação web desenhada pelo SLI para a consulta pública dos corpora paralelos do CLUVI. Com este trabalho, pretendemos contribuir para o avanço da investigação e desenvolvimento nas áreas da linguística de corpus e das tecnologias linguísticas da língua galega.

REFERÊNCIAS BIBLIOGRÁFICAS

- AGUIRRE MORENO, J.L., A. ÁLVAREZ LUGRÍS e X. GÓMEZ GUINOVART. 2002. “Etiquetario morfosintáctico del SLI para corpus de lengua gallega: aplicación al corpus paralelo TECTRA”. Em *Procesamiento del Lenguaje Natural*, 28, pp. 23-34.
- AGUIRRE MORENO, J.L., N. ANDIÓN e X. GÓMEZ GUINOVART. 2001. “Aspectos ortográficos, léxicos y morfosintácticos del etiquetado lingüístico de un corpus de informática en lengua gallega”. Em *Procesamiento del Lenguaje Natural*, 27, pp. 13-19.
- ÁLVAREZ, R., X.L. REGUEIRA e H. MONTEAGUDO. 1986. *Gramática galega*. Vigo: Galaxia.
- ÁLVAREZ, R. e X. XOVE. 2002. *Gramática da lingua galega*. Vigo: Galaxia.
- ÁLVAREZ LUGRÍS, A. 2001. *Estilística comparada da traducción: Proposta metodolóxica e aplicación práctica ó estudio do corpus TECTRA de traduccions do inglés ó galego*. Vigo: Universidade de Vigo.
- LEECH, G. e A. WILSON. 1996. *Recommendations for the Morphosyntactic Annotation of Corpora*. EAGLES Guidelines. Disponível em: <http://www.ilc.pi.cnr.it/eagles96/annotate/annotate.html>.
- MONACHINI, M. e N. CALZOLARI (coords.) 1996. *Synopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and Corpora*. EAGLES Guidelines. Disponível em: <http://www.ilc.pi.cnr.it/eagles96/morphsyn/morphsyn.html>.
- MONACHINI, M. e N. CALZOLARI. 1999. “Standardization in the Lexicon”. Em H. VAN HALTEREN (ed.), *Syntactic Wordclass Tagging*, pp. 149-174. Dordrecht: Kluwer.
- RAG/ILG. 1982. *Normas ortográficas e morfolóxicas do idioma galego*. RAG/ILG: Vigo (13ª edição revista: 1995).
- SAMPSON, G. 1995. *English for the Computer*. Oxford: Oxford University Press.
- TEUFEL, S. 1996. *ELM-DE: EAGLES Specifications for German Morphosyntax*. EAGLES Guidelines. Disponível em: http://www.ilc.pi.cnr.it/eagles96/elm_de/elm_de.html.

ANEXO 1: TECTRA (SECÇÃO INGLÊS)

A seguir apresenta-se uma lista com os títulos dos textos originais ingleses que fazem parte do corpus TECTRA, com o nome do autor/a, ano de publicação e tamanho em número de palavras.

Original inglês, autor/a (ano de publicação)	Tamanho
<i>The Pearl</i> , John Steinbeck (1945)	26.476
<i>Animal Farm</i> , George Orwell (1945)	30.533
<i>To the Lighthouse</i> , Virginia Woolf (1927)	70.836
<i>The Call of the Wild</i> , Jack London (1903)	31.960
<i>Extracts from Adam's Diary</i> , Mark Twain (1893)	4.596
<i>Eve's Diary</i> , Mark Twain (1906)	7.036
<i>Spanish Galicia</i> , Aubrey F.G. Bell (1922)	40.543
<i>The Golem</i> , Isaac B. Singer (1982)	12.834
<i>Nine Stories</i> , J.D. Salinger (1948)	55.917
<i>The Catcher in the Rye</i> , J.D. Salinger (1945)	75.329
<i>A Portrait of the Artist as a Young Man</i> , James Joyce (1916)	83.641
<i>Lord of the Flies</i> , William Golding (1954)	62.052
<i>The Third Man</i> , Graham Greene (1950)	31.793
<i>A Sentimental Journey</i> , Laurence Sterne (1768)	42.620
Total TECTRA – secção inglês	575.166

Quadro 17: TECTRA (secção inglês)

ANEXO 2: TECTRA (SECÇÃO GALEGO)

Neste segundo anexo recolhemos a lista dos títulos das traduções para o galego que fazem parte do corpus TECTRA, com o nome do tradutor/a, ano de publicação e tamanho em número de palavras.

Tradução galega, tradutor/a (ano de publicação)	Tamanho
<i>A perla</i> , Benigno F. Salgado (1990)	24.907
<i>A revolta dos animais</i> , X. Antón L. Dobao (1992)	26.215
<i>Cara ó faro</i> , Manuela Palacios & Xavier Castro (1993)	69.015
<i>A chamada da selva</i> , Gonzalo Navaza (1982)	29.053
<i>Retallos do diario de Adán</i> , Benigno F. Salgado (1992)	4.513
<i>Diario de Eva</i> , B. F. Salgado (1991)	6.771
<i>Galicia vista por un inglés</i> , X. M. Gómez Clemente (1994)	45.554
<i>O Golem</i> , Anxo Romero Louro (1989)	12.027
<i>Nove contos</i> , X. Antón L. Dobao (1994)	55.579
<i>O vixia no centeo</i> , X. Ramón F. Rodríguez (1990)	74.757
<i>Retrato do artista cando novo</i> , Vicente Araguas (1994)	82.398
<i>O señor das moscas</i> , X. M. Gómez Clemente (1993)	62.732
<i>O terceiro home</i> , M ^a Dolores M. Torres (1994)	30.833
<i>Unha viaxe sentimental</i> , Manuel Outeiriño (1992)	42.524
Total TECTRA – secção galego	551.878

Quadro 18: TECTRA (secção galego)

ANEXO 3: REFERÊNCIAS BIBLIOGRÁFICAS DO CORPUS TECTRA

Por último, indicamos as referências bibliográficas correspondentes aos romances (originais em inglês e traduções para o galego) incluídas no corpus paralelo TECTRA:

- BELL, A.F.G. 1922. *Spanish Galicia*. John Lane The Bodley Head, Londres.
- BELL, A.F.G. 1994. *Galicia vista por un inglés*. Galaxia, Vigo.
- GOLDING, W. 1954 (1962). *Lord of the Flies*. Faber & Faber, Londres.
- GOLDING, W. 1993. *O Señor das moscas*. Sotelo Blanco, Santiago de Compostela.
- GREENE, G. 1950 (1974, 3ª ed.). *The Third Man*. Heinemann, Londres.
- GREENE, G. 1994. *O terceiro home*. Galaxia, Vigo.
- JOYCE, J. 1916 (1986, 11ª ed.). *A Portrait of the Artist as a Young Man*. Grafton Books, Londres.
- JOYCE, J. 1994. *Retrato do artista cando novo*. Laivento, Santiago de Compostela.
- LONDON, J. 1903 (1975, 16ª ed.). *The Call of the Wild*. Heinemann, Londres.
- LONDON, J. 1982 (1983, 2ª ed.). *A chamada da selva*. Gerais, Vigo.
- ORWELL, G. 1945 (1987, 58ª ed.). *Animal Farm*. Penguin, Londres.
- ORWELL, G. 1992. *A revolta dos animais*. Positivas, Santiago de Compostela.
- SALINGER, J.D. 1948 (1986). *For Esmé, with Love and Squalor*. Penguin, Nova York.
- SALINGER, J.D. 1994. *Nove contos*. Sotelo Blanco, Santiago de Compostela.
- SALINGER, J.D. 1951. *The Catcher in the Rye*. Penguin, Londres.
- SALINGER, J.D. 1990 (1992, 4ª ed.). *O vixia no centeo*. Gerais, Vigo.
- SINGER, I. B. 1982. *The Golem*. Penguin, Londres.
- SINGER, I. B. 1989. *O Golem*. Gerais, Vigo.
- STEINBECK, J. 1945 (1986, 21ª ed.). *The Pearl*. Penguin, Nova York.
- STEINBECK, J. 1990 (1993, 5ª ed.). *A perla*. Galaxia, Vigo.
- STERNE, L. 1768 (1995). *A Sentimental Journey Through France and Italy*. Wordsworth, Hertfordshire.
- STERNE, L. 1992. *Unha viaxe sentimental por Francia e Italia*. Sotelo Blanco, Santiago de Compostela.
- TWAIN, M. 1893 (1993). *Extracts from Adam's Diary*. Courage Books, Filadélfia.
- TWAIN, M. 1992. *Retallos do diario de Adán*. Positivas, Santiago de Compostela.
- TWAIN, M. 1906 (1993). *Eve's Diary*. Courage Books, Filadélfia.
- TWAIN, M. 1991. *Diario de Eva*. Positivas, Santiago de Compostela.
- WOLF, V. 1927 (1977, 18ª ed.). *To the Lighthouse*. Grafton Books, Londres.
- WOLF, V. 1993. *Cara ó faro*. Sotelo Blanco, Santiago de Compostela.