CNIO - ANOTACIÓN SEMI-AUTOMÁTICA DE CORPUS CLÍNICO

Plan de impulso de las Tecnologías del Lenguaje

Carmen Torrijos Caruda

Nuria Aldama García

Octubre 2018

Este estudio ha sido realizado dentro del ámbito del Plan de Impulso de las Tecnologías del Lenguaje con financiación de la Secretaría de Estado para el Avance Digital, que no comparte necesariamente los contenidos expresados en el mismo. Dichos contenidos son responsabilidad exclusiva de sus autores.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

CNIO – ANOTACIÓN SEMI-AUTOMÁTICA DE CORPUS CLÍNICO

1 INTRODUCCIÓN

El Plan de Impulso de las Tecnologías del Lenguaje (Plan TL) tiene como objetivo fomentar el desarrollo del Procesamiento del Lenguaje Natural (PLN) y la Traducción Automática (TA) en lengua española y lenguas cooficiales. Para ello, el Plan TL define medidas que:

- Aumenten el número, calidad y disponibilidad de las infraestructuras lingüísticas en español y lenguas cooficiales.
- Impulsen la Industria del lenguaje fomentando la transferencia de conocimiento entre el sector investigador y la industria.
- Incorporen a la Administración como impulsor del sector de PLN.

Uno de los objetivos del proyecto es poner a disposición de la comunidad científica y la industria un corpus biomédico exhaustivo y con licencia abierta que permita ejecutar tareas de PLN sobre big data y replicar los experimentos. Este documento presenta los resultados del acuerdo entre anotadores en la fase de anotación manual, que comprende la segmentación de frases, la segmentación de formas y la anotación categorial de textos médicos en español. Consultar las guías de anotación para conocer los criterios de anotación manual en los tres niveles. Consultar el documento Metodología de anotación de textos biomédicos en español para conocer los detalles relativos a los perfiles del autor de la guía y de los anotadores del corpus.

2 DESCRIPCIÓN DEL PROCESO

Este informe recoge los datos de acuerdo entre anotadores del proceso de anotación manual y semiautomática del corpus clínico sobre el que trata el proyecto *CNIO – Anotación semi-automática de corpus clínico*.

El proyecto aquí presentado ha consistido en la anotación de un corpus clínico que contiene 1000 textos no estructurados relativos a casos clínicos. Se llevó a cabo una anotación semi-automática, es

decir, un 20% del corpus fue anotado manualmente por anotadores, mientras que el 80% restante fue anotado de manera automática con una versión implementada de la herramienta FreeLing3.1, que imita los criterios marcados por los anotadores.

La anotación se realizó en tres niveles: anotación de la segmentación de oraciones (split), anotación de la segmentación de formas (tokenización) y etiquetado según la categoría morfológica (part of speech 'POS'). Para llevar a cabo esta anotación se elaboraron tres guías, una por cada nivel, en las que quedan recogidos los criterios según los cuales se ha anotado manualmente el corpus.

El trabajo de anotación se distribuyó en tres fases. En la primera, los anotadores anotaron de forma paralela e individual el 10% del corpus de desarrollo (100 textos elegidos aleatoriamente), detectando al mismo tiempo casos excepcionales del dominio médico que debían ser incluidos en las guías de anotación (abreviaturas, unidades de medida, siglas, expresiones numéricas y alfanuméricas). Tras la anotación manual de ese 10% de desarrollo, se procedió a la extracción de los valores de acuerdo entre anotadores reflejados en la sección 3.1.

En la segunda fase del proceso se llevó a cabo la armonización de las anotaciones realizadas manualmente sobre el 10% de desarrollo, con el fin de solventar las discrepancias en aquellos casos en los que cada anotador había propuesto un análisis distinto en cualquiera de los tres niveles. De este modo se generó un *gold standard* de anotación de textos clínicos a partir de las anotaciones manuales, y se llevaron a cabo los ajustes necesarios en FreeLing3.1 para mejorar el acuerdo. Tras la armonización se calcularon de nuevo los valores de acuerdo entre el *gold standard* y FreeLing3.1 (ver sección 3.2).

En la fase final del proceso de anotación manual, uno de los anotadores anotó el 10% de validación del corpus (100 textos elegidos aleatoriamente y no contenidos en el 10% de desarrollo) según los criterios recogidos en las guías y las decisiones tomadas en el proceso de armonización. A su vez, se procedió a anotar el 10% de validación con la versión mejorada de FreeLing3.1 y se calculó el acuerdo entre la anotación manual y la anotación automática (ver sección 3.3). Alcanzados en el corpus de validación los valores mínimos requeridos de acuerdo entre anotadores para cada uno de los niveles, se procedió a la anotación automática del 100% del corpus.

Por último, el corpus anotado en su totalidad en formato FreeLing3.1 fue traducido al formato BRAT.

3 DESCRIPCIÓN CUANTITATIVA DEL CORPUS

El corpus en formato texto plano está compuesto por 1.000 textos no estructurados que reflejan casos clínicos. Cada texto médico equivale a un caso clínico y constituye un fichero de texto (.txt) de extensión variable.

El corpus cuenta con 64.865 oraciones, 353.124 palabras y 18.281 lemas diferentes. La ratio de palabras por oración es de 5,44. Las tablas 1 y 2 recogen las cien formas y los cien lemas más frecuentes del corpus (frecuencia absoluta y relativa).

	Forma	Frecuencia absoluta	Frecuencia relativa
1	de	26847	0.08
2	la	11644	0.03
3	y	11241	0.03
4	con	7584	0.02
5	en	7570	0.02
6	el	6114	0.02
7	se	5966	0.02
8	a	4917	0.01
9	una	4193	0.01
10	que	4172	0.01
11	del	3858	0.01
12	un	3377	0.01
13	por	3082	0.01
14	los	2322	0.01
15	paciente	2154	0.01
16	Se	1992	0.01
17	La	1789	0.01
18	En	1711	0.00
19	sin	1710	0.00

20	El	1681	0.00
21	las	1626	0.00
22	no	1545	0.00
23	tratamiento	1544	0.00
24	años	1470	0.00
25	al	1462	0.00
26	para	1337	0.00
27	mg	1301	0.00
28	fue	1238	0.00
29	su	1092	0.00
30	meses	881	0.00
31	como	829	0.00
32	cm	791	0.00
33	derecho	782	0.00
34	estudio	761	0.00
35	izquierdo	748	0.00
36	realizó	721	0.00
37	días	714	0.00
38	exploración	704	0.00
39	A	692	0.00
40	normal	673	0.00
41	dolor	662	0.00
42	día	632	0.00
43	lesión	622	0.00
44	evolución	620	0.00
45	antecedentes	616	0.00
46	diagnóstico	615	0.00
47	ni	611	0.00

48	e	600	0.00
49	dl	598	0.00
50	renal	581	0.00
51	lo	580	0.00
52	derecha	573	0.00
53	2	557	0.00
54	es	553	0.00
55	nivel	509	0.00
56	izquierda	508	0.00
57	tras	495	0.00
58	horas	484	0.00
59	No	482	0.00
60	abdominal	482	0.00
61	masa	453	0.00
62	dos	450	0.00
63	durante	449	0.00
64	3	448	0.00
65	hasta	445	0.00
66	presenta	440	0.00
67	después	430	0.00
68	1	429	0.00
69	normales	427	0.00
70	edad	421	0.00
71	cuadro	411	0.00
72	ingreso	410	0.00
73	Tras	410	0.00
74	realiza	409	0.00
75	le	401	0.00

76	era	394	0.00
77	más	392	0.00
78	mediante	391	0.00
79	desde	390	0.00
80	presencia	389	0.00
81	ml	388	0.00
82	o	385	0.00
83	células	384	0.00
84	presentaba	377	0.00
85	alta	375	0.00
86	lesiones	373	0.00
87	ojo	370	0.00
88	4	365	0.00
89	mm	365	0.00
90	TAC	364	0.00
91	g	341	0.00
92	6	340	0.00
93	l	338	0.00
94	control	336	0.00
95	clínica	331	0.00
96	5	327	0.00
97	fueron	318	0.00
98	semanas	318	0.00
99	ecografía	314	0.00
100	mostró	309	0.00

Tabla 1: Relación de las cien formas más frecuentes en el corpus.

	Lema	Frecuencia absoluta	Frecuencia relativa
1	de	26885	0.08
2	el	26107	0.07
3	y	11841	0.03
4	en	9281	0.03
5	uno	8083	0.02
6	se	7958	0.02
7	con	7810	0.02
8	a	4917	0.01
9	que	4172	0.01
10	de+el	3861	0.01
11	ser	3362	0.01
12	por	3173	0.01
13	paciente	2399	0.01
14	no	2027	0.01
15	realizar	1842	0.01
16	sin	1786	0.01
17	a+el	1748	0.00
18	presentar	1725	0.00
19	año	1714	0.00
20	tratamiento	1585	0.00
21	para	1392	0.00
22	día	1352	0.00
23	mg	1340	0.00
24	su	1287	0.00
25	2	1199	0.00
26	normal	1104	0.00
27	mes	1103	0.00

28	izquierdo	1059	0.00
29	lesión	1003	0.00
30	derecho	1001	0.00
31	como	964	0.00
32	tras	905	0.00
33	estudio	883	0.00
34	haber	848	0.00
35	cm	791	0.00
36	mostrar	790	0.00
37	3	788	0.00
38	exploración	779	0.00
39	dl	759	0.00
40	1	758	0.00
41	diagnóstico	735	0.00
42	antecedente	728	0.00
43	A	692	0.00
44	dolor	688	0.00
45	este	684	0.00
46	observar	659	0.00
47	durante	653	0.00
48	evolución	631	0.00
49	renal	629	0.00
50	nivel	627	0.00
51	ni	615	0.00
52	negativo	561	0.00
53	hora	546	0.00
54	abdominal	525	0.00
55	semana	525	0.00

56	4	523	0.00
57	masa	507	0.00
58	ojo	501	0.00
59	estar	484	0.00
60	alto	475	0.00
61	después	472	0.00
62	otro	463	0.00
63	6	462	0.00
64	hasta	457	0.00
65	decidir	455	0.00
66	ml	455	0.00
67	desde	454	0.00
68	varón	453	0.00
69	control	445	0.00
70	quirúrgico	440	0.00
71	ambos	436	0.00
72	ingreso	435	0.00
73	inferior	433	0.00
74	cuadro	425	0.00
75	edad	424	0.00
76	iniciar	423	0.00
77	5	419	0.00
78	mediante	416	0.00
79	le	403	0.00
80	derecha	402	0.00
81	clínico	399	0.00
82	encontrar	396	0.00
83	más	396	0.00

84	célula	395	0.00
85	presencia	392	0.00
86	dar	392	0.00
87	acudir	389	0.00
88	o	385	0.00
89	referir	385	0.00
90	alteración	372	0.00
91	tratar	368	0.00
92	mm	365	0.00
93	superior	364	0.00
94	TAC	364	0.00
95	ecografía	363	0.00
96	destacar	363	0.00
97	posterior	361	0.00
98	ante	360	0.00
99	apreciar	358	0.00
100	zona	357	0.00

Tabla 2. Relación de los cien lemas más frecuentes del corpus.

4 INFORME DE ACUERDO ENTRE ANOTADORES

4.1 FASE 1: ANOTACIÓN DEL 10% DE DESARROLLO.

En esta fase del proyecto se anotan los textos pertenecientes al 10% de desarrollo con la herramienta FreeLing3.1 adaptado al dominio médico. Paralelamente, los anotadores anotan a mano e individualmente el 10% de desarrollo. Este 10% de desarrollo está constituido por 100 textos clínicos escogidos aleatoriamente. La tabla 3 recoge los valores de acuerdo entre:

- anotador 1 y anotador 2
- anotador 1 y FreeLing3.1
- anotador 2 y FreeLing3.1.

	Split	Token	POS
A1 vs A2	**99.79**%	**99.97**%	98.85%
A1 vs FrL	99.37%	99.95**%**	98.47%
A2 vs FrL	99.58%	99.96%	**98.87%**
Acuerdo Mínimo Requerido	99%	98%	96%

Tabla 3. Porcentaje de acuerdo entre anotadores y FreeLing3.1 en el 10% de desarrollo.

El valor más alto de acuerdo entre anotadores en segmentación de oraciones se da entre el anotador 1 y el anotador 2 con un valor de discrepancia del 0.21%. En el caso de la segmentación por formas, el valor de acuerdo más elevado es alcanzado por la pareja anotador 1 y anotador 2, con una discrepancia de tan solo el 0.03%. Cabe destacar que este valor se encuentra a una y dos centésimas respectivamente por encima del resto de valores de acuerdo entre anotadores en segmentación de formas. El valor más alto de acuerdo para el etiquetado morfológico se da entre el anotador 2 y el anotador automático FreeLing3.1, que alcanzan un 98.87% de acuerdo.

4.2 FASE 2: ARMONIZACIÓN Y GOLD STANDARD.

Con el fin de contrastar y validar las anotaciones realizadas por el anotador 1 y anotador 2, se procedió a la armonización del 10% de desarrollo, es decir, a la resolución de aquellos casos en los que los anotadores habían propuesto distintas estrategias de split, tokenización o etiquetado morfológico.

3.2.1 Discrepancias en segmentación de oraciones entre anotadores

En el 10% de desarrollo tan solo se producen 3 discrepancias entre los anotadores en lo que se refiere a segmentación de oraciones. El 100% tiene que ver con la falta de separación entre dos oraciones cuando la primera de ellas termina con una abreviatura acabada en punto.

3.2.2 Discrepancias en tokenización entre anotadores

Los anotadores discrepan en la segmentación de formas en diez textos en el 10% de desarrollo. Estas discrepancias se producen debido a: segmentación de nombres propios (Ej: *Saint John of God*), segmentación de siglas (Ej: i.m.), adición de punto finalizador de oración a una sigla (Ej: C.).

3.2.3 Discrepancias en etiqueta morfológica entre anotadores

En la tabla 5 se recogen los casos de discrepancia en etiqueta morfológica que se repiten más de una vez entre anotador 1 y anotador 2, siguiendo patrón recogido en la tabla 4.

'[Etiqueta propuesta por anotador 1 – forma : etiqueta propuesta por anotador 2 – forma]'

Tabla 4.Patrón de discrepancia entre anotadores.

	Discrepancia	Frecuencia absoluta	Frecuencia relativa (%)
1	'[VMG0000+P00CN000 : VMG0000+PP3CN000]'	62	14,73
2	'[AQ0FS0 : NCFS000]'	24	5,7
3	'[NC00000 : NP00000]'	23	5,46
4	'[AQ0MS0 : NCMS000]'	20	4,75
5	'[VMN0000+P00CN000 : VMN0000+PP3CN000]'	16	3,8
6	'[NC00000 : NCFS000]'	15	3,56
7	'[NCMS000 : AQ0MS0]'	13	3,09
8	'[NCMS000 : AQ0CS0]'	12	2,85
9	'[NCFS000 : AQ0FS0]'	10	2,38
10	'[NC00000 : VAIP3S0]'	9	2,14
11	'[NCCS000 : AQ0CS0]'	9	2,14
12	'[NCMS000 : NCFS000]'	9	2,14
13	'[Z : NC00000]'	8	1,9
14	'[AQ0CS0 : NCCS000]'	7	1,66
15	'[NCFS000 : NP00000]'	7	1,66
16	'[VAIP3S0 : NC00000]'	7	1,66
17	'[NC00000 : SPS00]'	7	1,66
18	'[NCMP000 : AQ0MP0]'	5	1,19
19	'[NP00000 : NCFS000]'	5	1,19
20	'[NC00000 : NCMS000]'	5	1,19
21	'[NP00000 : NCMS000]'	5	1,19

22	'[Z : Z]'	4	0,95
23	'[NCMS000 : AQ0FS0]'	4	0,95
24	'[AQ0CS0 : AQ0MS0]'	4	0,95
25	'[NC00000 : NC00000]'	4	0,95
26	'[NCCS000 : NCFS000]'	4	0,95
27	'[NCFS000 : VMP00SF]'	3	0,71
28	'[PR0CN000 : CS]'	3	0,71
29	'[AQ0MP0 : NCMP000]'	3	0,71
30	'[NCFS000 : NCMS000]'	3	0,71
31	'[P03CN000 : P00CN000]'	3	0,71
32	'[SPS00 : NCFS000]'	2	0,48
33	'[NCFS000 : NCCS000]'	2	0,48
34	'[NCFS000 : SPS00]'	2	0,48
35	'[AQ0CS0 : AQ0CP0]'	2	0,48
36	'[NCFS000 : VMIP3S0]'	2	0,48
37	'[VMIP3S0 : AQ0FS0]'	2	0,48
38	'[AQ0MS0 : PI0MS000]'	2	0,48
39	'[NCMP000 : AQ0CP0]'	2	0,48
40	'[NCMS000 : NCCS000]'	2	0,48
41	'[RG : AQ0MS0]'	2	0,48
42	'[RG : VMIP1S0]'	2	0,48
43	'[RG : NCMS000]'	2	0,48
44	'[AQ0FP0 : NCFP000]'	2	0,48
45	'[AQ0MS0 : SPS00]'	2	0,48
46	'[VMP00SF : NCFS000]'	2	0,48
47	'[NCCS000 : NCMS000]'	2	0,48
48	'[SPS00 : NP00000]'	2	0,48
49	'[VMP00SM : NCMS000]'	2	0,48

50	'[PI0MS000 : AQ0MS0]'	2	0,48
51	'[NCMS000 : AQMS0]'	1	0,24
52	'[CC : NCFS000]'	1	0,24
53	'[AQ0CP0 : AQ0FP0]'	1	0,24
54	'[SPS00 : S]'	1	0,24
55	'[VMP00PF : NCFP000]'	1	0,24
56	'[RG : NC00000]'	1	0,24
57	'[NCMS000 : VMSP3S0]'	1	0,24
58	'[NC0000 : NC00000]'	1	0,24
59	'[VMII3S0 : NP00000]'	1	0,24
60	'[NCMS000 : NC00000]'	1	0,24
61	'[NCMS000 : VMN0000]'	1	0,24
62	'[NCFP000 : NC00000]'	1	0,24
63	'[RG : NCFS000]'	1	0,24
64	'[NCFS000 : NCFP000]'	1	0,24
65	'[NCMS000 : VMIP3S0]'	1	0,24
66	'[SPS00 : AQ0MS0]'	1	0,24
67	'[NCMP000 : NCMS000]'	1	0,24
68	'[VAN0000+P00CN000 : VAN0000+PP3CN000]'	1	0,24
69	'[NCFS000 : NCFN000]'	1	0,24
70	'[AQ0MS0 : AQMS00]'	1	0,24
71	'[NCMS000 : NP00000]'	1	0,24
72	'[NCFS000 : AQ0CS0]'	1	0,24
73	'[NP00000 : RG]'	1	0,24
74	'[NCMS000 : NCMP000]'	1	0,24
75	'[AQ0FS0 : VMIP3S0]'	1	0,24
76	'[AQ0CS0 : AQCS0]'	1	0,24
77	'[NCFP000 : NP00000]'	1	0,24

78	'[SPS000 : SPS00]'	1	0,24
79	'[NCFN000 : NCFS000]'	1	0,24
80	'[CS : PR0CN000]'	1	0,24
81	'[AQ00000 : NCMS000]'	1	0,24
82	'[VMP00SF : AQ0FS0]'	1	0,24
83	'[NCMN000 : NCMP000]'	1	0,24
84	'[AQ0MS : NCMS000]'	1	0,24
85	'[NP00000 : NP00000]'	1	0,24
86	'[AQ0MS0 : AQ0MP0]'	1	0,24
87	'[AQ0CP0 : NCMP000]'	1	0,24
88	'[AQ0CS0 : NCFS000]'	1	0,24
89	'[RG : PP3CNO00]'	1	0,24
90	'[RG : AQ0CS0]'	1	0,24
91	'[NCFS000 : RG]'	1	0,24
92	'[NC00000 : Z]'	1	0,24
93	'[NCFS000 : NFCS000]'	1	0,24
94	'[VMG0000+P00CN000+PP3CSD00 : VMG0000+PP3CN000+PP3CSD00]'	1	0,24
95	'[NCCS000 : NCCS00]'	1	0,24
96	'[PI0FS000 : DI0FS0]'	1	0,24
97	'[AQ0MS0 : VMP00SM]'	1	0,24
98	'[NC0CS0 : AQ0CS0]'	1	0,24
99	'[NCCS0 : AQ0CS0]'	1	0,24
100	'[AQ0CS0 : AQ0FS0]'	1	0,24
101	'[NP00000 : NCFP000]'	1	0,24
102	'[NP00000 : DA0FS0]'	1	0,24
103	'[NP00000 : DD0FS0]'	1	0,24
104	'[NP00000 : NC00000]'	1	0,24

105	'[PP3CN000 : P00CN000]'	1	0,24
106	'[AQ0MP0 : NC00000]'	1	0,24
107	'[VMIP3S0 : NC00000]'	1	0,24
108	'[Z : SPS00]'	1	0,24
109	'[NCMS000 : VMIP1S0]'	1	0,24
110	'[PI0MS000 : PI0FS000]'	1	0,24
111	'[AQ0FS0 : AQ0CS0]'	1	0,24
112	'[Z : NCFS000]'	1	0,24
113	'[NCCS000 : NP00000]'	1	0,24
114	'[VMP00PM : AQ0MS0]'	1	0,24
115	'[VMIP1S0 : VMIS3S0]'	1	0,24
116	'[AQ0CS0 : NCMS000]'	1	0,24
117	'[DA0MP0 : NP00000]'	1	0,24
118	'[NCMS000 : AQ0CP0]'	1	0,24
119	'[AQ0MS0 : NCMP000]'	1	0,24
120	'[AQ0CS0 : NCMP000]'	1	0,24
TOTAL		421	100

Tabla 5. Relación de discrepancias en POS entre anotadores

Nota: la coindidencia en las etiquetas con discrepancia en POS son discrepancias de tokenización

En el 10% de desarrollo del corpus se encuentran 421 discrepancias en etiqueta morfológica. Los tres casos más frecuentes de discrepancia entre anotadores responden a errores en la anotación, el etiquetado de 'se' (P00CN000, PP3CN000, P03CN000) y palabras ambiguas que pueden comportarse tanto como sustantivos como como adjetivos (Ej. 'derecho', 'donante', 'paciente' o 'negativa').

Para solventar los casos de discrepancia, se observó el contexto de aparición de la forma sobre la cual había que tomar una decisión y, si cualquiera de las etiquetas morfológicas propuestas por los anotadores tenía cabida en el citado contexto, se tenía entonces en cuenta la etiqueta propuesta por FreeLing3.1 como método desambiguador, siempre y cuando la etiqueta propuesta por FreeLing3.1 coincidiese con alguna de las dos etiquetas propuestas por los anotadores. La ilustración 1 muestra este proceso.

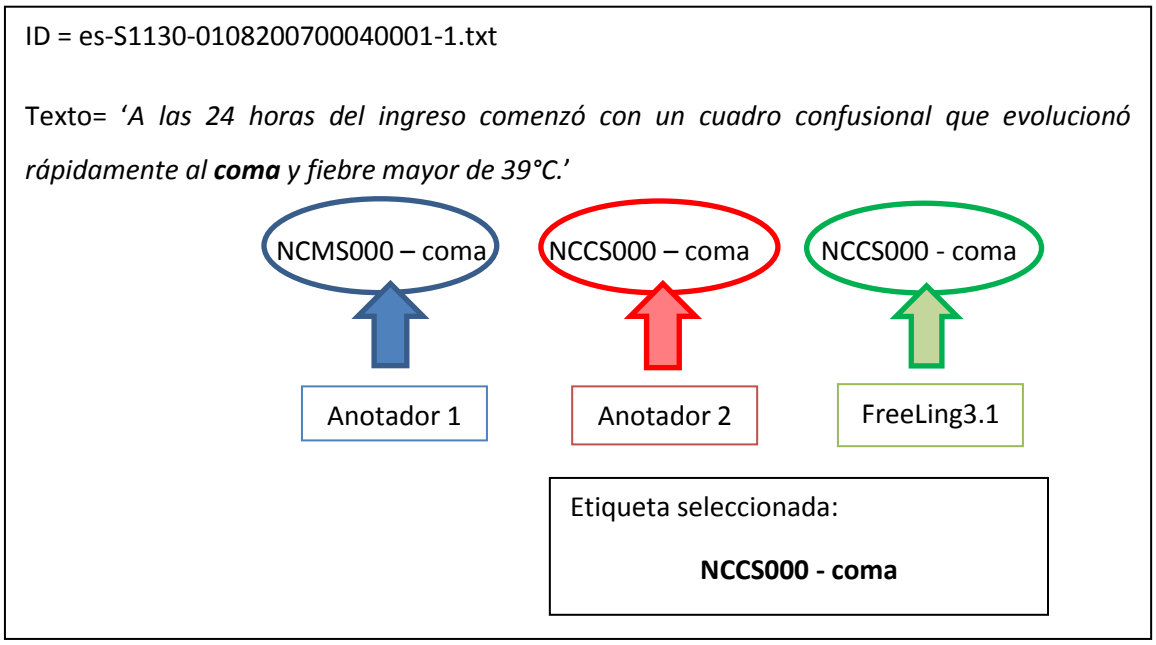

Ilustración 1: protocolo de desambiguación

Además, cabe destacar que hay seis textos que en los que no ha habido discrepancias entre los anotadores en ninguno de los niveles a anotar.

Una vez solventadas las discrepancias, y por tanto, construido el *gold standard*, se extrajeron de nuevo los valores de acuerdo entre FreeLing3.1 y el *gold standard*. Dichos valores se encuentran recogidos en la tabla 6.

	Split	**Token**	**POS**
GS vs FrL	99.37%	99.95%	98.36%
Acuerdo Mínimo Requerido	99%	98%	96%

Tabla 6. Porcentaje de acuerdo entre el gold standard y FreeLing3.1 sobre 10% de desarrollo

Los casos que generan discrepancia entre el *gold standard* y FreeLing3.1 quedan expuestos a continuación.

3.2.4 Discrepancias en segmentación de oraciones entre gold standard y FreeLing3.1

Las discrepancias a nivel de segmentación oracional entre FreeLing3.1 y la anotación armonizada del 10% de desarrollo afectan a siete casos clínicos. Estas discrepancias se generan debido a siglas en

posición final de frase acabadas en punto (Ej: 'estadío Ib._Posteriormente', 'dcha._La' 'mmol/L._Recibía'), y segmentación de siglas que incluyen puntos en varios tokens (Ej: i.m.).

3.2.5 Discrepancias en tokenización entre gold standard y FreeLing3.1

Las discrepancias en tokenización afectan a 16 textos. Estas discrepancias se deben a:

- Segmentación de siglas que incluyen puntos en varios tokens (Ej: i.m.),
- Abreviaturas acabadas en punto en posición final de frase (Ej: 'estadío/ Ib. _Posteriormente'),
- Segmentación de nombres propios (Ej: 'Saint_John/ of/ God', 'Unidad_de_Cirugía_Bucal/ y/ Maxilofacial'),
- Adición de punto finalizador de frase siglas no terminadas en punto (Ej: 'Mitomicina/C_.'),
- Segmentación del punto que finaliza abreviaturas en tokens separados (Ej: a/.).

3.2.6 Discrepancias en etiquetado morfológico entre gold standard y FreeLing3.1

En la tabla 7 se recogen los casos más frecuentes de discrepancia en etiqueta morfológica entre el gold standard y FreeLing3.1.

	Discrepancia	Frecuencia absoluta	Frecuencia relativa (%)
1	'[AQ0FS0 : NCFS000]'	78	12,96
2	'[AQ0MS0 : NCMS000]'	66	10,96
3	'[VMG0000+P00CN000 : VMG0000+PP3CN000]'	62	10,3
4	'[NCFS000 : AQ0FS0]'	33	5,48
5	'[Z : NC00000]'	26	4,32
6	'[NCFS000 : NP00000]'	24	3,99
7	'[NC00000 : NP00000]'	23	3,82
8	'[NCMS000 : NP00000]'	20	3,32
9	'[VMN0000+P00CN000 : VMN0000+PP3CN000]'	16	2,66
10	'[AQ0CS0 : NCMS000]'	15	2,49

11	'[NC00000 : NCFS000]'	13	2,16
12	'[NCMS000 : AQ0MS0]'	11	1,83
13	'[PI0MS000 : AQ0MS0]'	9	1,5
14	'[VMP00SF : NCFS000]'	8	1,33
15	'[AQ0CS0 : NCCS000]'	7	1,16
16	'[NCMP000 : AQ0MP0]'	7	1,16
17	'[PI0FS000 : AQ0FS0]'	7	1,16
18	'[NP00000 : NC00000]'	6	1
19	'[NCCS000 : AQ0CS0]'	6	1
20	'[NP00000 : NP00000]'	5	0,83
21	'[NC00000 : NC00000]'	5	0,83
22	'[AQ0MP0 : NCMP000]'	5	0,83
23	'[NCMS000 : NCFS000]'	5	0,83
24	'[AQ0CP0 : NCMP000]'	4	0,66
25	'[NCFS000 : NCMS000]'	4	0,66
26	'[VMIP3S0 : AQ0FS0]'	4	0,66
27	'[AQ0CP0 : AQ0CS0]'	4	0,66
28	'[RG : VMIP1S0]'	4	0,66
29	'[VMIP3S0 : NCFS000]'	4	0,66
30	'[NC00000 : SPS00]'	4	0,66
31	'[AQ0CS0 : NC00000]'	4	0,66
32	'[PR0CN000 : CS]'	3	0,5
33	'[RG : NC00000]'	3	0,5
34	'[RG : AQ0MS0]'	3	0,5
35	'[NC00000 : NCMS000]'	3	0,5
36	'[NC00000 : VAIP3S0]'	3	0,5
37	'[NCMP000 : NP00000]'	3	0,5
38	'[AQ0MS0 : SPS00]'	3	0,5

39	'[P03CN000 : P00CN000]'	3	0,5
40	'[NCMP000 : AQ0CP0]'	3	0,5
41	'[AQ0CS0 : NCFS000]'	3	0,5
42	'[NP00000 : NCFS000]'	2	0,33
43	'[NCMS000 : AQ0CS0]'	2	0,33
44	'[NCCS000 : AQ0000]'	2	0,33
45	'[AQ0FS0 : NC00000]'	2	0,33
46	'[NCFN000 : NCFS000]'	2	0,33
47	'[RG : NP00000]'	2	0,33
48	'[AQ0MS0 : NP00000]'	2	0,33
49	'[NCFS000 : NC00000]'	2	0,33
50	'[RG : NCMS000]'	2	0,33
51	'[AQ0FP0 : NCFP000]'	2	0,33
52	'[AQ0CP0 : NCFP000]'	2	0,33
53	'[AO0FS0 : Z]'	2	0,33
54	'[SPS00 : NP00000]'	2	0,33
55	'[AQ0MP0 : NCCP000]'	2	0,33
56	'[AQ0MS0 : NCFS000]'	2	0,33
57	'[NCMS000 : VMIP1S0]'	2	0,33
58	'[NCFP000 : AQ0FP0]'	1	0,17
59	'[NCFP000 : VMP00PF]'	1	0,17
60	'[DA0NS0 : PP3CNA00]'	1	0,17
61	'[NCFS000 : VMP00SF]'	1	0,17
62	'[NCMS000 : NC00000]'	1	0,17
63	'[NCMS000 : VMSP3S0]'	1	0,17
64	'[VMII3S0 : NP00000]'	1	0,17
65	'[VSN0000 : NCMS000]'	1	0,17
66	'[VMP00SM : VMP00MS]'	1	0,17

67	'[RG : NCFS000]'	1	0,17
68	'[VMIP3S0 : NCMS000]'	1	0,17
69	'[NCMS000 : NCMP000]'	1	0,17
70	'[VAN0000+P00CN000 : VAN0000+PP3CN000]'	1	0,17
71	'[SPS00 : RG]'	1	0,17
72	'[RG : AQ0CS0]'	1	0,17
73	'[AQ0CS0 : NP00000]'	1	0,17
74	'[RG : VMSP1S0]'	1	0,17
75	'[AQ0MS0 : RG]'	1	0,17
76	'[AQ0CS : AQ0CS0]'	1	0,17
77	'[NCFP000 : NP00000]'	1	0,17
78	'[SPS000 : SPS00]'	1	0,17
79	'[CS : PR0CN000]'	1	0,17
80	'[NCMN000 : NCMS000]'	1	0,17
81	'[Z : Z]'	1	0,17
82	'[AQ0FS0 : VMIP3S0]'	1	0,17
83	'[NCMS000 : RG]'	1	0,17
84	'[NCFS000 : SPS00]'	1	0,17
85	'[AQ0MS0 : NCMP000]'	1	0,17
86	'[VMN0000+PP3CNA00 : NC00000]'	1	0,17
87	'[Z : NP00000]'	1	0,17
88	'[AQ0MP0 : NC00000]'	1	0,17
89	'[NCFS000 : VMIP3S0]'	1	0,17
90	'[RG : PP3CNO00]'	1	0,17
91	'[NCFS000 : NCCS000]'	1	0,17
92	'[VMG0000+P00CN000+PP3CSD00 : VMG0000+PP3CN000+PP3CSD00]'	1	0,17
93	'[PI0FS000 : DI0FS0]'	1	0,17

94	'[VMP00PM : AQ0MP0]'	1	0,17
95	'[VMP00SM : AQ0MS0]'	1	0,17
96	'[NCFS000 : VMM03S0]'	1	0,17
97	'[NCCP000 : AQ0CP0]'	1	0,17
98	'[NCMS000 : VMP00SM]'	1	0,17
99	'[DA0FS0 : NP00000]'	1	0,17
100	'[DD0FS0 : NP00000]'	1	0,17
101	'[PP3CN000 : P00CN000]'	1	0,17
102	'[VMIP3S0 : NC00000]'	1	0,17
103	'[VMP00SM : NCMS000]'	1	0,17
104	'[Z : NCFS000]'	1	0,17
105	'[VMIS3S0 : VMIP1S0]'	1	0,17
106	'[DA0MP0 : NP00000]'	1	0,17
107	'[NCMS000 : AQ0CP0]'	1	0,17
108	'[AQ0CS0 : NCMP000]'	1	0,17
TOTAL		602	100

Tabla 7. Relación de discrepancias en POS entre gold standard y FreeLing3.1

El acuerdo entre FreeLing3.1 y el *gold standard* contiene 602 discrepancias en etiquetado morfológico. Las discrepancias más frecuentes son aquellas relacionadas con adjetivos y sustantivos y el uso de 'se' en construcciones en las que el clítico va unido a un gerundio.

4.3 FASE 3: ANOTACIÓN DEL 10% DE VALIDACIÓN.

Con el fin de validar la anotación manual realizada por los anotadores, así como las mejoras implementadas en FreeLing3.1, se extrajo un 10% del corpus (100 textos aleatorios que no estuvieran contenidos en el 10% de desarrollo). Este 10% de validación fue anotado manualmente por un solo anotador siguiendo los criterios empleados para la obtención del *gold standard*, criterios recogidos en las guías de anotación. A su vez, este mismo 10% fue anotado de manera automática por freeLing3.1 en su versión final. Tras la anotación manual y automática, se procedió a calcular los valores de acuerdo entre anotador y FreeLing3.1. Dichos valores quedan recogidos en la siguiente tabla.

	Split	Token	POS
GS vs FrL	99.52%	99.97%	98.71%
Acuerdo Mínimo Requerido	99%	98%	96%

Tabla 8.Porcentaje de acuerdo sobre el 10% de validación

3.2.4 Discrepancias en segmentación de oraciones en el 10% de validación

Las discrepancias a nivel de segmentación oracional entre FreeLing3.1 y el 10% de validación afectan a 7 textos del total. Dichas discrepancias tienen que ver principalmente con unidades de medida, siglas o abreviaturas en posición final de oración a las que se les añade un punto (Ej: '°_C. La', 'U/L. El', 'etc. El'). En estos casos, FreeLing3.1 interpreta el punto como finalizador de token y no como finalizador de oración, y por tanto no segmenta la oración correctamente.

3.2.5 Discrepancias en tokenización en el 10% de validación

Las discrepancias en tokenización afectan a 13 textos. Estas discrepancias se deben a segmentación de siglas que incluyen abreviaturas, siglas y unidades de medida acabadas en punto en posición final de frase (Ej: 'L_.', 'cc_.'), segmentación de nombres propios (Ej: 'T./Millin'), adición de tokens a un nombre propio (Ej: 'Servicio_de_Cirugía_Buco-Maxilofacial_de_la_Facultad_de_Odontología_de_Piracicaba'), o segmentación de siglas que contienen puntos (Ej: 'i/./m/.').

3.2.6 Discrepancias en etiquetado morfológico en el 10% de validación

La tabla 9 contiene los casos en los que la anotación del *gold standard* discrepa de la anotación proporcionada por FreeLing3.1 sobre el 10% de validación.

	Discrepancias	Frecuencia absoluta	Frecuencia relativa (%)
1	'[VMG0000+P00CN000 : VMG0000+PP3CN000]'	96	19,59
2	'[AQ0FS0 : NCFS000]'	82	16,73
3	'[NCFS000 : AQ0FS0]'	39	7,96
4	'[AQ0MS0 : NCMS000]'	39	7,96

5	'[VMN0000+P00CN000 : VMN0000+PP3CN000]'	21	4,29
6	'[Z : NC00000]'	18	3,67
7	'[AQ0MP0 : NCMP000]'	12	2,45
8	'[VMP00SF : NCFS000]'	10	2,04
9	'[P03CN000 : P00CN000]'	9	1,84
10	'[PI3FS000 : AQ0FS0]'	7	1,43
11	'[PI0FS000 : AQ0FS0]'	7	1,43
12	'[NP00000 : NP00000]'	6	1,22
13	'[RG : AQ0MS0]'	6	1,22
14	'[AQ0FS00 : AQ0FS0]'	6	1,22
15	'[AQ0CS : AQ0CS0]'	6	1,22
16	'[NCMS000 : NC00000]'	5	1,02
17	'[NCMS000 : NP00000]'	5	1,02
18	'[AQ0MS0 : SPS00]'	5	1,02
19	'[NCFS000 : NP00000]'	5	1,02
20	'[RG : VMIP1S0]'	4	0,82
21	'[NCMS000 : AQ0MS0]'	4	0,82
22	'[AQ0FP0 : NCFP000]'	4	0,82
23	'[NCFS000 : NC00000]'	4	0,82
24	'[NC00000 : NC00000]'	3	0,61
25	'[AQ0CS0 : NCMS000]'	3	0,61
26	'[NCFS000 : SPS00]'	2	0,41
27	'[AQ0CS0 : NCCS000]'	2	0,41
28	'[NCMP000 : AQ0CP0]'	2	0,41
29	'[AQ0MS0 : NCMP000]'	2	0,41
30	'[RG : NCFS000]'	2	0,41
31	'[NC00000 : NP00000]'	2	0,41
32	'[NCMS000 : NCFS000]'	2	0,41

33	'[VMP00SM : NCMS000]'	2	0,41
34	'[AQ0MS0 : NC00000]'	2	0,41
35	'[AQ0CS0 : NC00000]'	2	0,41
36	'[SPS00 : NC00000]'	2	0,41
37	'[NCMN000 : NCMN000]'	2	0,41
38	'[VMIS3S0 : NCMS000]'	2	0,41
39	'[VMG0000+P03CN000 : VMG0000+PP3CN000]'	2	0,41
40	'[AQ0CS0 : VMN0000]'	2	0,41
41	'[VMP00SM : VMG0000]'	1	0,2
42	'[NCCS000 : AQ0CS0]'	1	0,2
43	'[AQ0MS0 : NCCS000]'	1	0,2
44	'[NCMS000 : NCMP000]'	1	0,2
45	'[NCCS000 : AQ0CP0]'	1	0,2
46	'[CS : PR0CN000]'	1	0,2
47	'[NC00000 : SPS00]'	1	0,2
48	'[VMG0000+P00CN000+PP3CSD00 : VMG0000+PP3CN000+PP3CSD00]'	1	0,2
49	'[NCFS000 : PI0FS000]'	1	0,2
50	'[NC00000 : VMIP3S0]'	1	0,2
51	'[AQ0FS0 : NP00000]'	1	0,2
52	'[VMN0000+P00CN000+PP3CSD00 : VMN0000+PP3CN000+PP3CSD00]'	1	0,2
53	'[PI0MP000 : AQ0MP0]'	1	0,2
54	'[PI3MS000 : AQ0MS0]'	1	0,2
55	'[NCCS000 : NP00000]'	1	0,2
56	'[NCCN000 : AQ0CN0]'	1	0,2
57	'[Z : NP00000]'	1	0,2
58	'[AQ0FP0 : NC00000]'	1	0,2
59	'[AQ0CS0 : NP00000]'	1	0,2

60	'[NP00000 : NC00000]'	1	0,2
61	'[DA0MS0 : NP00000]'	1	0,2
62	'[AQ0MS00 : AQ0MS0]'	1	0,2
63	'[AQ0CS0 : VMSP1S0]'	1	0,2
64	'[NCFS000 : AQ0FP0]'	1	0,2
65	'[VMP00SF : AQ0FS0]'	1	0,2
66	'[NCMS000 : VMP00SM]'	1	0,2
67	'[NCCP000 : AQ0CP0]'	1	0,2
68	'[AQ0MP0 : NCCP000]'	1	0,2
69	'[RG : NC00000]'	1	0,2
70	'[SPS00 : NP00000]'	1	0,2
71	'[DA0MP0 : NP00000]'	1	0,2
72	'[VMN0000+P03CN000 : VMN0000+PP3CN000]'	1	0,2
73	'[NCMS000 : VMIP1S0]'	1	0,2
74	'[PI0MS000 : AQ0MS0]'	1	0,2
75	'[NCMP000 : NCFP000]'	1	0,2
76	'[NCFN000 : NC00000]'	1	0,2
77	'[NC00000 : AQ0CS0]'	1	0,2
78	'[VAN0000+P00CN000 : VAN0000+PP3CN000]'	1	0,2
79	'[NCFS000 : NCFS000]'	1	0,2
80	'[NCFS000 : VMIP3S0]'	1	0,2
81	'[NCFS000 : I]'	1	0,2
82	'[AQ0MS0 : VMIP1S0]'	1	0,2
83	'[NCFS000 : NCMS000]'	1	0,2
84	'[NCFP000 : NP00000]'	1	0,2
85	'[AQ0FS0 : VMIP3S0]'	1	0,2
86	'[PR0CN000 : CS]'	1	0,2
87	'[AQ0CS0 : NCFS000]'	1	0,2

88	'[PR3CN000 : CS]'	1	0,2
89	'[VMP00PF : NCFP000]'	1	0,2
90	'[NCMP000 : VMSP2S0]'	1	0,2
91	'[NCFN000 : NCFS000]'	1	0,2
92	'[DA0FS0 : NP00000]'	1	0,2
93	'[VMIP3S0 : NCFS000]'	1	0,2
94	'[RG : AQ0CN0]'	1	0,2
TOTAL		490	100

Tabla 9. Relación de discrepancias entre gold standard y FreeLing3.1 sobre 10% validación.

Las discrepancias en etiqueta morfológica entre la anotación del *gold standard* y la de FreeLing3.1 suman un total de 490. Las discrepancias más frecuentes se producen en torno al etiquetado de 'se' cuando este va precedido de un gerundio. A esta tipología de discrepancia le siguen, como en casos anteriores, aquellas discrepancias que surgen en aquellos casos en los que un sustantivo funciona como un modificador nominal o cuando un adjetivo se ve sustantivado.

5 CONCLUSIONES

Todos los procesos anteriores confluyen en un corpus de 1000 textos procedentes de la narrativa clínica anotados en tres niveles lingüísticos: segmentación de oraciones, segmentación de formas o tokenización y etiquetado morfológico (POS). Los procesos de interanotator agreement proporcionan, por un lado, la garantía de que los anotadores han seguido un proceso riguroso de anotación, validado por los altos valores de acuerdo entre ellos, y por otro lado la seguridad de que la herramienta de anotación automática (FreeLing3.1) presenta una discrepancia mínima con la anotación manual, lo que garantiza la calidad de su adaptación al dominio médico.

Este corpus anotado permite la aplicación de capas posteriores para la detección de entidades médicas, la implementación de modelos de aprendizaje automático y el desarrollo de diversas soluciones por parte del Plan Nacional de Tecnologías del Lenguaje.

6 REFERENCIAS

[1] M. Oronoz, K. Gojenola, A. Pérez, A.D. de Ilarraza and A.Casillas. "On the creation of a clinical gold standard corpus in Spanish: Mining adverse drug reactions", in Journal of Biomedical Informatics, 56, 2015, pp. 318-332.

[2] S.V. Pakhomov, A. Coden and C.G. Chute. "Developing a corpus of clinical notes manually annotated for part-of-speech", in International Journal of Medical Informatics, 75, 2006, pp. 418-429.

[3] L. Campillos, L. Deléger, C. Grouin, T. Hamon, A.L Ligozat and A. Névéol. "A French clinical corpus with comprehensive semantic annotations: development of the Medical Entity and Relation LIMSI annotated Text corpus (MERLOT)", in Language Resources & Evaluation, 52, 2018, pp. 571-601.

7 GLOSARIO DE SIGLAS Y ACRÓNIMOS

A1	Anotador 1
A2	Anotador 2
CNIO	Centro Nacional de Investigaciones Oncológicas
Ej	Ejemplo
FrL	FreeLing3.1
GS	Gold standard
POS	Part of Speech
VS	Versus
Plan TL	Plan Nacional de Tecnologías del Lenguaje
PLN	Procesamiento de Lenguaje Natural
TA	Traducción Automática