



UNIVERSITAT_{DE}
BARCELONA

Facultat de Filologia
i Comunicació

Grau de Lingüística

Treball de Fi de Grau

Curs 2020-2021

Sobre l' anotació de les relacions argumentals en debats

Criteris per l' anotació de les relacions entre arguments en debats de competició en català

Nom de l'estudiant:

Daniel Muñoz Toledano

Tutores:

Montserrat Nofre i Mariona Taulé

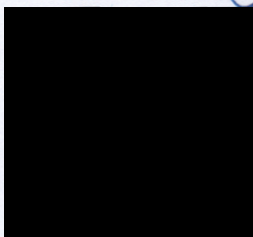
Barcelona, 2021



Amb aquest escrit declaro que sóc l'autor/autora original d'aquest treball i que no he emprat per a la seva elaboració cap altra font, incloses fonts d'Internet i altres mitjans electrònics, a part de les indicades. En el treball he assenyalat com a tals totes les citacions, literals o de contingut, que procedeixen d'altres obres. Tinc coneixement que d'altra manera, i segons el que s'indica a l'article 18, del capítol 5 de les Normes reguladores de l'avaluació i de la qualificació dels aprenentatges de la UB, l'avaluació comporta la qualificació de "Suspens".

Barcelona, a 10 de juny del 2021

Signatura:



AGRAÏMENTS

Mirant la portada o la referència d'una investigació, treball o article, podem saber el nom de totes les persones que hi han participat. Però n'hi ha moltes altres sense la seva participació tangencial i indirecta de les quals no hagués estat possible tirar-los endavant; persones que ens envolten i amb les quals compartim la nostra vida. Aquest apartat està dedicat a aquestes persones imprescindibles, sense les quals no hagués pogut fer aquest treball.

En primer lloc, a les meves tutores del TFG, Montserrat i Mariona, que des del primer moment m'han ofert la seva experiència i saviesa i m'han guiat per tal que aquest treball arribi a bon port. Estic molt agraït d'haver pogut treballar i compartir temps amb vosaltres. Moltes gràcies per tota l'ajuda que m'heu prestat.

En segon lloc, al meu company David, perquè el treball d'anotació no hagués estat igual sense tu, amb tots els moments que hem compartit al llarg d'aquest curs i carrera. Et desitjo el millor en tot i pots comptar amb mi per a qualsevol cosa.

A continuació, agrair a l'investigador Ramon Ruiz Dolz la informació tècnica de l'eina que està desenvolupant per a la seva tesi i del programa que van fer servir per transcriure els debats, per respondre totes les meves preguntes.

Tot seguit, agrair a la meva família tot el que han fet per mi: a la meva mare Fuensanta per estar amb mi en tot moment i donar-me escalf quan el necessitava; a la meva germana Maria per compartir temps, passatemps i somriures; i al meu pare Juan per totes les xerrades i totes les estones que hem compartit junts.

A continuació, els meus amics de tota la vida: Alberto, Guillem, Víctor, Pol, Alberto, Anna i Maria. Gràcies per compartir les poques estones que hem pogut, fer-me riure i poder desconectar una estona. I també per tots els consells que m'heu donat durant aquest any.

Per últim, agrair a l'Helena, per compartir somnis i alegries, però també per estar en els moments durs donant-me suport. Perquè sense tu aquest any estrany hagués estat molt més difícil i perquè has fet la meva experiència universitària molt més dolça. Moltes gràcies per estar al meu costat.

Índex

Declaració d'autoria

1	Introducció	1
1.1	Objectius del treball	1
1.2	Estat de la qüestió	2
1.3	Treballs previs	4
2	Metodologia	8
2.1	Obtenció de les dades	8
2.2	Eines d' anotació	9
2.2.1	OVA	9
2.2.2	Alternativa d' anotació: dades estructurades	10
2.3	Transcripció	12
2.3.1	Varietats lingüístiques	12
2.3.2	Puntuació	12
2.3.3	Noms propis i paraules estrangeres	13
2.4	Segmentació	13
2.4.1	Definició de les ADUs	14
2.4.2	Criteris de segmentació en ADUs	14
2.4.2.1	Coordinació i juxtaposició	14
2.4.2.2	Oracions subordinades	16
2.4.2.3	Exemples i metàfores	18
2.4.2.4	Altres criteris	19
2.5	Anotació	20
2.5.1	Identificador de l'ADU	20
2.5.2	Part del debat	21
2.5.3	Número d'argument	21
2.5.4	Posicionament (<i>stance</i>)	21
2.5.5	Identificador de la relació	22
2.5.6	Tipus de relació	22

3	Resultats	23
3.1	Proves d'acord entre anotadors	23
3.1.1	Acord en la transcripció	23
3.1.2	Acord en la segmentació i l'anotació	24
3.2	Resultats de l'anotació del corpus	25
3.3	Representació en grafs	27
3.4	Discussió	28
4	Conclusions	29
	Bibliografia	31
	Índex de taules	33
	Índex de figures	33
	Apèndix A <i>Script</i> Bash per a la identificació d'ADUs aïllades	34
	Apèndix B Codi de l'<i>script</i> per generar grafs	35

Resum

El present treball tracta sobre el procés d' anotació del corpus Vives Debate. Per a la recopilació del corpus s'han fet servir enregistraments dels debats de la Lliga de Debats Universitària, organitzada per la Xarxa Vives.

L' anotació del corpus s'ha dividit en tres tasques: transcripció, segmentació i anotació. Per cadascuna d'aquestes tasques s'han establert diversos criteris per tal que l' anotació sigui homogènia entre els anotadors. Aquests criteris estan detallats i són l' objectiu del present treball.

El corpus s'ha creat amb l' objectiu de servir com entrenament per a eines d' anàlisi automàtica de les relacions argumentals en debats. Aquestes eines, actualment, només es troben entrenades en anglès. Per tant, el corpus Vives Debate és el primer del seu tipus en català.

Paraules clau: Anotació, corpus, discurs argumentatiu, relacions argumentals.

This paper addresses the annotation process of the Vives Debate corpus. For the compilation of this corpus, recordings of the debates of the Lliga de Debats Universitària, organized by Xarxa Vives, have been used.

The annotation was divided into three tasks: transcription, segmentation and annotation. For each of these tasks, several criteria have been established so that the annotation is homogeneous among the annotators. These criteria are explained and are the focus of the present work.

The corpus has been created to serve as a training for tools for automatic analysis of argumentative relations in debates. These tools are currently only trained in english. Therefore, the Vives Debate corpus is the first of its kind in Catalan.

Keywords: Annotation, corpus, argumentative speech, argumental relations.

1. INTRODUCCIÓ

1.1. Objectius del treball

El present treball descriu el procés d'anotació d'un corpus de debats i, de forma específica, els criteris d'anotació de les relacions entre unitats discursives argumentals. El corpus amb el que s'ha treballat és Vives Debate, un corpus integrat per debats universitaris de la Lliga de Debat Universitària. El corpus s'ha creat com a base documental per a la recerca de Ramon Ruiz Dolz, investigador del *Valencian Research Institute for Artificial Intelligence* (VRAIN) de la Universitat Politècnica de València, amb la col·laboració de la Xarxa Vives, organitzadora de la Lliga de Debat Universitària, que ha donat accés als enregistraments audiovisuals dels debats.

La part del corpus Vives Debate anotada fins ara, a la que fa referència aquest treball, consta de 12 debats, 11 corresponents a l'edició 2019 i 1 corresponent a l'edició 2018 de la Lliga de Debats Universitària. Actualment el corpus es segueix anotant en una segona fase, seguint els criteris ja establerts. En aquesta fase s'anotaran uns altres 18 arxius fins arribar a un total de 30 arxius.

El procés d'anotació del corpus Vives Debate consta de tres tasques diferenciades:

1. Revisió i correcció de la transcripció automàtica del debat.
2. Segmentació en unitats argumentals discursives (ADUs).
3. Anotació de les relacions entre els arguments, classificats en tres grans categories:
 - Inferència, quan una proposició dona una raó que condueix a l'acceptació de l'altra.
 - Conflicte, quan dues proposicions es contradiuen en algun dels seus aspectes.
 - Reformulació, quan una proposició en reformula una altra.

En un principi estava previst dur a terme l'anotació amb el programa OVA,¹ seguint el model proposat per treballs anteriors (Visser *et al.*, 2020). En estudis anteriors on s'havia utilitzat aquesta eina s'analitzaven textos d'una extensió força més reduïda que els que componen el nostre corpus; en intentar utilitzar el programa amb el nostre corpus, ens vam adonar que el seu funcionament amb una quantitat major de dades no era òptim i provocava problemes tècnics que dificultaven una anotació correcta i eficient. Per tot això es va optar per una eina més

¹<http://ova.arg-tech.org/>

genèrica que permetés una major flexibilitat en l’ anotació de textos d’ una mida relativament gran i que, a la vegada, funcionés en plataformes diferents i fos versàtil i suficientment coneguda pels anotadors, per poder solucionar eventuais errors que poguessin aparèixer. Per tot això es va considerar que la millor alternativa era crear una estructura d’ anotació partint d’ un full de càlcul amb algunes opcions predefinides, amb la qual cosa els textos anotats s’ emmagatzemaven en forma de dades estructurades. Un cop solucionats els problemes tècnics inicials, vam procedir a establir els criteris d’ anotació, basant-nos en treballs previs (Visser *et al.*, 2020) i en la casuística que anava apareixent durant la pròpia anotació.

Aquest treball descriu el procés d’ anotació, presa de decisions i establiment dels criteris del corpus Vives Debate. Cal remarcar que és el primer recurs d’ aquest tipus que es crea en llengua catalana i, també, el primer que s’ anota amb relacions argumentals. A més, l’ extensió final del corpus superarà la d’ altres projectes similars en altres llengües.

1.2. Estat de la qüestió

L’ argumentació ha estat objecte d’ estudi i debat des dels orígens de l’ estudi del discurs a la retòrica grega, al segle V a. de C. Des del punt de vista de la retòrica, l’ argumentació són les estratègies utilitzades per un orador amb la intenció de convèncer el seu interlocutor. En la actualitat, el marc teòric de Anscombe i Ducrot (Escandell Vidal, 2013) estableix que l’ acte d’ argumentar implica presentar un enunciat, o un conjunt d’ enunciats, perquè s’ admeti un segon enunciat, que és la conclusió. Seguint aquest model podem afirmar que l’ argumentació és un conjunt d’ enunciats interrelacionats i presentats de forma estructurada.

La mineria d’ argumentació, *argument mining* en anglès, consisteix en una sèrie de tècniques desenvolupades en el marc de la lingüística computacional i de corpus, mitjançant les quals és possible detectar automàticament la presència d’ arguments dintre d’ un discurs i obtenir-ne informació lingüística.

Històricament, la mineria d’ argumentació s’ ha dut a terme aplicant regles basades en estudis lingüístics, però la incorporació de les noves tecnologies de la computació, com el *deep learning*, ha permès l’ extracció de dades d’ una manera més automatitzada (Chakrabarty *et al.*, 2020). Aquest tipus de tractament de la informació permet treballar amb un major volum de dades, en comparació amb la manera més tradicional de fer-ho, a través de regles lingüístiques, i de forma molt més ràpida. Els mètodes basats en *deep learning* utilitzen un corpus anotat per

entrenar el sistema d'una manera autònoma o semiguiada. Alguns treballs que segueixen aquest model són el ja esmentat de Chakrabarty *et al.* (2020) o el d'Esteve (2020).

AMPERSAND (Chakrabarty *et al.*, 2020) proposa un mètode computacional d'extracció d'arguments per discussions en fòrums en línia. El mètode es basa en la predicció de les relacions a partir de la informació contextual extreta d'un model lingüístic entrenat i de la parametrització de les relacions de la teoria estructural de la retòrica. El treball d'Esteve (2020) es centra en la mineria d'argumentació i forma part de la tasca MultiStanceCat (Taulé *et al.*, 2018).

La informació obtinguda pels sistemes de mineria d'argumentació s'han utilitzat per fer anàlisi de sentiments, anàlisi d'opinions i detecció de fal·làcies en el discurs, entre d'altres fenòmens que es poden investigar amb aquest mètode. Un exemple d'aquests usos de la mineria d'argumentació es l'estudi de Han i Kando (2019), on es realitza una extracció dels arguments per detectar l'opinió dels interlocutors. El model que fan servir està basat en un model de *word embedding*, que, explicat de forma resumida, és una tècnica que fa servir un model del llenguatge representat per vectors de números reals.

Un debat es defineix com “una qüestió entre dues o més persones de posicions contraposades, les quals es troben subjectes a les regles metòdiques d'intervenció i d'exposició de llurs arguments, sota la supervisió d'un conductor”(IEC, 2021). Sota aquesta premissa del debat com un acte de convèncer a un tercer s'han dut a terme competicions de debats a mode d'esport intel·lectual. Com tots els esports, tenen un seguit de reglaments per poder participar i puntuar les competicions.

Existeixen molts formats de debats (Wikipedia, 2021), dels quals els més destacats són:

- El debat Karl Popper, que es basa en l'intercanvi d'idees entre dues postures enfrontades, amb el suport de dades empíriques. El format es divideix en 10 intervencions: 6 sense interrupcions i 4 torns de preguntes entre els equips.
- El debat Lincoln-Douglas, que és un format de debat on solament hi ha dos debatents. Les argumentacions es basen en la moral i la lògica. Consta de 7 torns: 2 de discurs argumentats, 3 torns de refutació i 2 de preguntes entre els debatents.
- El debat estil Oxford, que és un debat competitiu entre dos equips que defensen postures enfrontades. En aquest estil de debat el públic assistent intervé fent preguntes als debatents i valorant l'argumentació.

- El debat a l'estil del Parlament Britànic, que és un tipus de debat entre equips on participen tots els membres seguint un ordre preestablert. El debat es realitza en dos torns: la introducció, on debaten els dos primers debatents dels equips intercalant-se, i les conclusions, on participen la resta de membres dels equips intercalant-se.

Les regles que fa servir la Xarxa Vives (Xarxa Vives, 2019), l'organisme que organitza els debats del corpus utilitzat en aquest treball, es basen en el model del Parlament Britànic, però amb modificacions. Les regles principals són:

- La llengua utilitzada és el català.
- El tema està prefixat i és el mateix per tots els debats de l'edició.
- Els equips, formats per estudiants universitaris, han de defensar la seva posició assignada mitjançant un sorteig.
- El debat consta de tres parts: introducció, dos torns de refutació i conclusions.
- El debat té una durada de 40 minuts i cada part té un temps assignat.
- Els equips poden fer interpellacions durant el torn de refutació de l'altre equip.

Les regles exposades condicionen com és el corpus Vives Debate. En primer lloc, com que la part real de debat de cada arxiu és de aproximadament de 40 minuts, el corpus és extens. Per una altra banda, l'existència de quatre torns dels intervenció implica que les quatre parts del debat es relacionin entre elles. A més, la possibilitat de les interrupcions per part dels participants durant els torns de refutació provoca que les argumentacions s'aturin i es reprenguin després de contestar la pregunta. També provoca que hi hagi ADUs d'un equip dins del torn de refutació d'un altre equip.

Per últim, les argumentacions poden comptar amb el suport d'elements extralingüístics contextuals com ara diapositives, imatges o documents, que no es poden recuperar en el corpus. En conseqüència, es poden trobar ADUs amb expressions díctiques i altres continguts que en facin referència però no hi ha correlat en el corpus.

1.3. Treballs previs

Un corpus semblant en l'objecte, els debats, i amb unes característiques similars, a excepció de l'extensió, és el corpus US2016 (Visser *et al.*, 2020), un corpus anotat dels debats televisius que

van mantenir els candidats a la presidència dels Estat Units d'Amèrica de l'any 2016, Hillary Clinton i Donald Trump, i les reaccions del públic a la xarxa social Reddit, un lloc web amb subfòrums on els usuaris poden escriure posts amb material audiovisual i enllaços de tercers.

Subcorpus	Tokens
Corpus debats televisius	58 900
Corpus Reddit	39 099
Total	97 999

Taula 1: Dades del US2016 (Visser *et al.*, 2020)

El sistema d' anotació del corpus US2016 s'ha utilitzat com a model per a l' anotació del corpus Vives Debate. Aquest sistema fa servir com anotació tres etiquetes: *Inference relation*, quan hi ha una relació directa entre proposicions; *Conflict relation*, quan existeix una incompatibilitat directa entre dues proposicions; i *Rephrase relation*, quan una proposició en refà una altra. A més d'aquestes etiquetes, l'US2016 també reconstrueix el contingut proposicional d'algunes ADUs.

L' anotació del corpus US2016 es va dur a terme amb l'eina OVA, que ja s'ha esmentat abans. Els principals trets d'aquesta eina són la interfície gràfica d' anotació i la possibilitat d'exportar els resultats en format imatge i en format de dades json. A més, permet publicar el corpus anotat.²

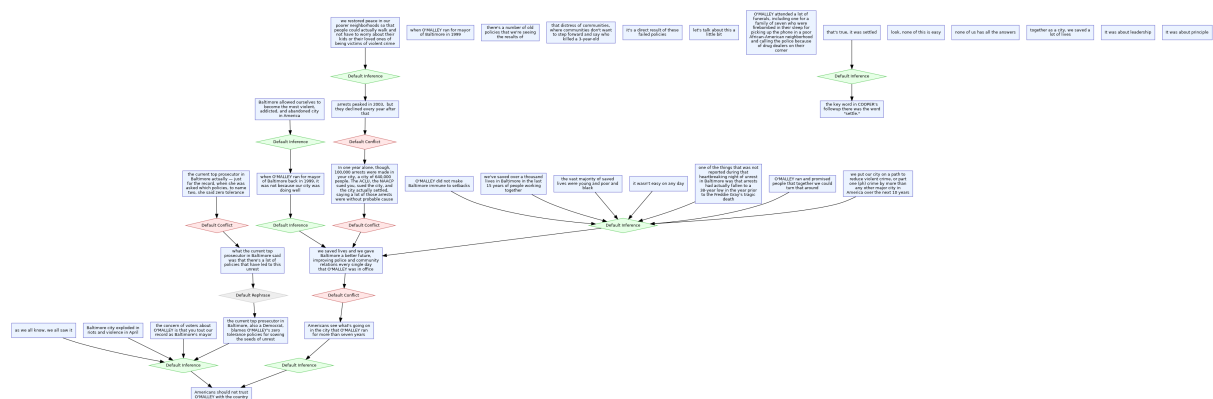


Figura 1: Mostra de l' anotació US2016 amb l'eina OVA

Un altre corpus amb una tipologia similar, un corpus de debats, és el recollit per Mićović (2014)

²analytics.arg-tech.org/overview.php?c=US2016

per a la seva tesi doctoral. Es tracta d'un corpus bilingüe en castellà i en serbi, integrat per quatre subcorpus:

- Transcripcions de tres debats electorals espanyols i dos debats electorals serbis, amb les seves corresponents traduccions.
- Respostes d'experts en comunicació política espanyola als qüestionaris proposats per l'autora.
- Entrevistes a dos moderadors de debats: Manuel Campo Vidal i Zoran Stanojevic, aquest últim versió original i traduït al castellà.
- Recopilació dels enregistraments audiovisuals.

El primer subcorpus recull els debats cara a cara en tots dos països en un moment cronològic similar, com es pot veure en la Taula 2.

	Debat 1	Debat 2	Debat 3
Espanya	2008	2008	2011
Sèrbia	2008	2012	

Taula 2: Anys dels debats (Mićović, 2014)

La principal diferència entre el corpus Mićović (2014) i el corpus Vives Debate rau en el tipus de dades que els integren. El corpus de Mićović (2014) conté un total de cinc debats, mentre que el Vives Debate en conté dotze (a la primera fase de l'anotació).

Una altra diferència és el tipus de debat. El corpus de Mićović (2014) conté debats entre dos polítics que defensen el seu posicionament ideològic, mentre que els debats de Vives Debate són entre equips universitaris que han de defensar un posicionament assignat per sorteig i, per tant, han de preparar-se prèviament tant els arguments a favor com els contraris.

A més, els debats de Vives Debate es guien per unes regles concretes i tenen uns jutges que qualifiquen la qualitat de l'argumentació. Per contra, els debats entre polítics estan menys regulats i el seu objectiu no és aconseguir una qualificació d'uns jutges imparcials, sinó persuadir els electors per aconseguir el seu vot.

L'última diferència es refereix al tractament i anàlisi del corpus. En el corpus Mićović (2014) es mantenen les transcripcions senceres dels debats per fer una anàlisi de les estratègies comuni-

catives i comparar-les. En canvi, el corpus Vives Debate es fa un treball de segmentació prèvia a l'anotació de les relacions argumentatives.

Un últim exemple d'un corpus de debats és el compilat per Wells (2018) a partir de les transcripcions dels debats entre els candidats de les eleccions presidencials de 2016 als Estats Units. És a dir, el contingut coincideix amb el corpus de Visser (2020). L'objectiu del treball és establir els criteris per a la detecció de les fal·làcies dins dels discurs polític, concretament de la fal·lacia de l'home de palla. El corpus es va anotar amb l'objectiu d'extreure patrons lingüístics i els elements lèxics i sintàctics que permetin la detecció automàtica de la fal·lacia. L'anotació va consistir a distingir les proposicions del debat entre les associades a la posició del debatent i les que l'oponent associa de manera fal·laç a la posició del debatent.

El corpus de Wells (2018) es diferencia de Vives Debate en diversos punts: la llengua dels debats (anglès), la quantitat de dades (és menor en el cas de Wells), els participants dels debats i la motivació del corpus que, en el cas de Wells, és la mineria de fal·làcies.

Com es pot observar a partir d'aquest exemples, el corpus Vives Debate té unes característiques que el diferencien de la resta: un corpus extens, en llengua catalana, anotat amb relacions discursives argumentals i creat amb l'objectiu de ser utilitzat com recurs d'entrenament i avaluació en sistemes de mineria d'arguments.

2. METODOLOGIA

2.1. Obtenció de les dades

Sigles	Universitat
UA	Universitat d'Alacant
UAB	Universitat Autònoma de Barcelona
UAO-CEU	Universitat Abat Oliba CEU
UB	Universitat de Barcelona
UdG	Universitat de Girona
UdL	Universitat de Lleida
UIB	Universitat de les Illes Balears
UJI	Universitat Jaume I
UMH	Universitat Miguel Hernández
UOC	Universitat Oberta de Catalunya
UPF	Universitat Pompeu Fàbra
UPV	Universitat Politècnica de València
URL	Universitat Ramón Llull
UV	Universitat de València

Taula 3: Correspondència entre les sigles i les universitats dels equips debatents

El corpus inicial de la primera fase de l'anotació constava de 12 vídeos, d'una durada aproximada d'uns 50 minuts cadascun, i la transcripció automàtica. Cada vídeo correspon a un debat de la Lliga de Debat Universitària

En la Taula 4 es pot veure en detall la durada de cada debat, que en total sumen 9 hores i 55 minuts de vídeos.

A la Taula 4 es detalla el contingut dels arxius que componen el corpus, en nombre de *tokens*, *types* i el càlcul de TTR (*Type Token Ratio*), indicativa de la riquesa lèxica del text. En total el corpus conté 77 807 *tokens*, que suposen una mitjana de 6 483,92 *tokens* per debat.

Debat	<i>Vídeos</i>	<i>Textos</i>		
	Duració	Tokens	Types	TTR
Debat UAB-UA	00:49:15	7 229	1 411	0,20
Debat UAB-UdG	00:50:01	7 365	1 482	0,20
Debat UB-UdL	00:50:49	5 271	1 193	0,23
Debat UB-UOC	00:49:28	5 978	1 300	0,22
Debat UdG-UIB	00:49:54	6 278	1 360	0,22
Debat UdL-UIB	00:49:21	5 486	1 330	0,24
Debat UJI-UPV	00:53:11	6 200	1 461	0,24
Debat UMH-UAO-CEU	00:48:14	6 865	1 341	0,20
Debat UPF-UJI	00:50:46	6 391	1 486	0,23
Debat UPF-UPV	00:51:43	7 085	1 530	0,22
Debat URL-UMH	00:43:36	7 189	1 428	0,20
Debat UV-UJI	00:50:28	6 470	1 524	0,24
Total				
	<i>Suma</i>	09:54:30	77 807	
	<i>Mitjana</i>	00:49:32	6 483,92	1 403,83 0,22

Taula 4: Durada dels vídeos i dades lexicomètriques de les transcripcions de la primera fase d' anotació del corpus Vives Debate

2.2. Eines d' anotació

2.2.1. OVA

Com ja hem explicat anteriorment (apartat 1.1), en un primer moment es va plantejar dur a terme l' anotació utilitzant el programari OVA (Janier *et al.*, 2014), un programa especialitzat per a l' anotació d' arguments mitjançant una interfície visual (Figura 2).



Figura 2: Exemple de la interfície d' anotació del programa OVA

Com podem observar, aquest programa permet seleccionar directament fragments del text introduït a la columna esquerra (1). Els segments seleccionats (ADUs) es creen en l'àrea central de treball (2) on s'anotaran les relacions entre segments a partir dels requadres d'il·locució (3). Durant el procés inicial d'anotació amb el programa OVA ens van sorgir alguns problemes, fonamentalment deguts a dues causes:

- Sincronització incorrecta dels nodes creats, que feia que nodes esborrats no s'esborressin del arxiu final.
- Extensió del corpus, que provocava que l'anotació amb la interfície fos complicada i es quedés bloquejada en moltes situacions.

Aquests inconvenients relantitzaven l'anotació i podien ser font de nombrosos errors, per la qual cosa vam considerar més eficient utilitzant un altre mètode d'anotació, amb una eina que ens permetés emmagatzemar els resultats en forma de dades estructurades.

2.2.2. Alternativa d'anotació: dades estructurades

Per substituir el programa OVA es va decidir utilitzar un programari de fulls de càlcul, com per exemple el Microsoft Excel o el Libreoffice Calc, per la seva capacitat de gestionar un volum de dades relativament gran, suficient per aquesta tasca, per la divisió en cel·les que permet

introduir les dades del corpus de manera senzilla i, per tant, solucionar els errors de etiquetatge d'una manera més simple. I, per últim, per la possibilitat de convertir-ho a format CSV per un posterior processament, si s'escau.

	A	B	C	D	E	F	G
1	ID (Chronological)	TYPE (PART + PERSON)	ARGUMENT NUMBER	TEAM STANCE	ADU	RELATED ID	ARGUMENTAL RELATION TYPE
2		1 INTRO		FAVOUR	viure a una societat on tinga la llibertat per a decidir sobre el meu cos.		
3		2 INTRO		FAVOUR	una societat on tinga el dret a decidir quan vull i quan no vull ser mare.	1 MA	
4		3 INTRO		FAVOUR	una societat que sembla utòpica.	1 RA	
5		4 INTRO		FAVOUR	però és la societat a la qual podem arribar.	1 CA	
6		5 INTRO		FAVOUR	una societat que ja ha aconseguit grans avanços com és la llei de l'avortament.	1 RA	
7		6 INTRO		FAVOUR	una societat que garanteix a les dones tindre la llibertat per a decidir quan no volen ser mare.	2 MA	
8		7 INTRO		FAVOUR	famílies pot garantir quan volen ser-ho.	6 MA	
9		8 INTRO		FAVOUR	una societat que • no ha donat compte de que existeix un problema i • lluita contra la clandestinitat.		
10		9 INTRO		FAVOUR	pot lluitar contra les malalties i contra el turisme reproductiu.		

Figura 3: Exemple d'un debat anotat en un full de càlcul

En la Figura 3 es pot veure com s'ha dut a terme l'anotació dels debats, a partir d'un format de columnes que es corresponen amb cada ítem a anotar.

L'anotació es divideix en 7 columnes:

1. ID: l'identificador de l'ADU.
2. TYPE: indica la posició del debat.
3. ARGUMENT NUMBER: indica a quin argument fan referència les ADUs.
4. TEAM STANCE: indica el posicionament de l'equip al qual pertany l'ADU.
5. ADU: el segment de text que integra l'ADU.
6. RELATED ID: indica l'identificador (columna ID) de les ADUs amb les que es relaciona l'ADU que s'està anotant.
7. ARGUMENTAL RELATION TYPE: indica el tipus de relació que s'estableix entre l'ADU que s'està anotant i l'ADU o ADUs amb les que es relaciona.

Per facilitar l'anotació i evitar errors, tant de etiquetatge com de format, a les columnes TYPE, TEAM STANCE i ARGUMENTAL RELATION TYPE s'han limitat les dades que es poden introduir mitjançant l'opció de validació de dades. La columna ID s'ha programat per generar automàticament una numeració consecutiva.

2.3. Transcripció

Entenem com a transcripció el procés pel qual un discurs oral és representat mitjançant un sistema de signes escrits. Normalment aquestes transcripcions utilitzen el sistema ortogràfic de l'idioma concret o un sistema de representació fonètica (per exemple l'IPA). Aquestes transcripcions es poden fer automàticament mitjançant un programa de STT (de l'anglès *Speech To Text*) que reconeix la parla i la transcriu.

Per a la compilació d'aquest corpus s'ha fet servir la segona opció de transcripció, un programa de transcripció automàtica, perquè la durada dels debats dificulta fer-ho de manera manual en un termini de temps relativament curt. Així doncs, atès que la transcripció pròpiament dita es duu a terme de manera automàtica, la tasca que anomenem “Transcripció” consisteix a revisar les transcripcions fetes pel programa a partir dels vídeos dels debats, esmenar possibles errors i normalitzar la transcripció de determinades formes.

Les transcripcions automàtiques es van obtenir amb el programa MLLP (2020), un programari Universitat Politècnica de València. Aquest programa utilitza models probabilístics basats en corpus de transcripcions realitzades de manera manual.

2.3.1. Varietats lingüístiques

El principal problema en aquesta etapa rau en les variacions lingüístiques, sobretot diatòpiques, presents en el corpus, motivades perquè en els debats hi participen universitats de tot el territori de parla catalana. En el procés de normalització dels textos s'ha intentat, sempre que ha estat possible, mantenir les variacions presents en la parla dels debatents. Podem veure'n en una intervenció de la equip de l'UIB, com l'article salat o les desinències verbals mallorquines:

- (1) i darrerament els majors errors sortiran a partir del que se coneixerà com la guerra contra el terror i l'invasió d'iraq que esdevindrà en una llarga sèrie d'aconteixements cadascun pitjor que s'anterior i un d'aquest aconteixements l'uniram en el nostre segon argument que és un argument sociològic que hem anomenat de sa reactància social (UdL-UIB 2018)

2.3.2. Puntuació

La transcripció generada pel programa no conté punts de puntuació ni símbols ni caràcters especials, a excepció del guionet i de l'apòstrof, per raons ortogràfiques. En el següent exemple

veiem com es manté el text sencer en minúscules i no s'afegeix cap signe de puntuació. Els únics signes que es apareixen són el guionet i el apòstrof.

- (2) perquè no tenen un incentiu per fer-ho el que fem és que hi hagi empreses i agències privades que s'aprofitin (UAB-UA 2019)

2.3.3. Noms propis i paraules estrangeres

Durant la revisió dels textos es van trobar diverses mencions a noms propis estrangers que el programa no reconeixia. Per aquests supòsits es va establir el criteri de corregir la transcripció sempre que el anotador conegués el nom propi en qüestió i, en cas contrari, transcriure'l adequant-lo als sons escoltats.

- (3) a. protegir a la gamyi de la discriminació que va patir (UB-UOC 2019)
- b. mercaderia de kadja x hedman i aquí queda palesa aquesta (UdG-UIB 2019)
- c. que passe ucraïna com demostren l'isaki moto doctora en dret i bioètica (URL-UMH 2019)

En alguns casos concrets de dubtes apareguts durant la transcripció, es va consultar el llibre d'estil de la Corporació Catalana de Mitjans Audiovisuals³ i el Diccionari de la llengua catalana de l'Institut d'Estudis Catalans.⁴

2.4. Segmentació

La fase de segmentació fa referència a la divisió del discurs en unitats discretes (*Argumentative Discursive Units*, ADUs), amb la finalitat de poder anotar les relacions argumentals que estableixen entre elles. En aquest procés s'ha hagut de consensuar què es considera una unitat segmentable i establir una sèrie de criteris, detallats en els següents apartats, per poder segmentar casos concrets que poden generar dubtes.

³<https://esadir.cat/>

⁴<https://dlc.iec.cat/>

2.4.1. Definició de les ADUs

Un debat, com hem explicat a l'apartat 1.2, està compost per diverses idees i arguments que s'estructuren a través de diferents enunciats al llarg del discurs. Això fa necessari que s'hagi de delimitar l'objecte de segmentació, per aconseguir que hi hagi homogeneïtat en els fragments delimitats pels anotadors, i que no hi hagi discrepàncies rellevants. L'estàndard escollit per aquest corpus, a proposta dels investigadors del VRAIN, ha estat l'*Argumentative Discursive Unit* establerta a Peldszus i Stede (2013), utilitzada també en el corpus US2016 (Visser *et al.*, 2020).

Les ADUs es defineixen com “la unitat mínima per a què un argument tingui significat propi i rellevància per al discurs”. Aquesta definició bàsica acota l'objecte d'estudi suficientment per definir un conjunt de criteris lògics i eficients de segmentació que ajudin a l'acord entre els anotadors davant de possibles dubtes o vacil·lacions. Aquest criteris s'expliquen en els següents apartats.

2.4.2. Criteris de segmentació en ADUs

En aquesta secció es descriuran tots els criteris establerts per la segmentació de les ADUs en diversos casos concrets que s'han trobat al corpus Vives Debate.

2.4.2.1. Coordinació i juxtaposició

La coordinació és el procés mitjançant el qual es fa una construcció amb dos o més elements que es troben al mateix nivell jeràrquic. Aquesta coordinació pot ser de tres tipus: copulativa, adversativa o disjuntiva.

A efectes de la segmentació, la coordinació dependrà dels elements que integren la construcció final. Es distingeixen les coordinacions d'elements aïllats (categories o sintagmes), i les coordinacions entre oracions.

En els exemples 4.a i 4.b veiem coordinació entre elements i en els exemples 5.a i 5.b una coordinació copulativa entre oracions.

(4) a. això perjudica a [les dones] i [els fills] (URL-UMH 2019)

b. en el sistema neoliberal [masclista] i [heteropatriarcal] (UV-UJI 2019)

- (5) a. [podem sortejar la prohibició legislativa al nostre país] i [aprofitar-nos de la misèria] (UV-UJI 2019)
- b. [la gestació subrogada és un fet] i [la seua prohibició coarta la meua generositat] (UAB-UA 2019)

Pel que fa a la coordinació d'elements, els elements coordinats formen part de la mateixa ADU, perquè es considera que els elements de la llista estan involucrats dins de la mateixa argumentació. No obstant això, si a la fase d'anotació es veu que una altra ADU fa referència a un element concret de la coordinació, els elements coordinats es segmentaran en ADUs independents.

Per altra banda, i a diferència de les coordinacions d'elements, les coordinacions d'oracions sempre es segmentaran en ADUs independents. La segona part de la coordinació mantindrà la conjunció, perquè pot servir d'ajuda a l'hora de establir el tipus de relació argumental que estableix amb l'ADU amb la qual es coordina.

- (6) a. ADU 1 podem sortejar la prohibició legislativa al nostre país
ADU 2 i aprofitar-nos de la misèria
- b. ADU 1 la gestació subrogada és un fet
ADU 2 i la seua prohibició coarta la meua generositat

Per últim, existeix un altre procés amb el qual podem unir constituents: la juxtaposició. La característica de la juxtaposició és la absència de nexa que relacioni els constituents de l'oració. A efectes de la segmentació, els elements i les oracions juxtaposades es tractaran de la mateixa manera que en els casos de segmentació.

- (7) a. i és que quan una dona està esdevenint [un atuell un forn un recipient] (UV-UJI 2019)
- b. la gestació subrogada [és un fet] [és una realitat] (UAB-UA 2019)

2.4.2.2. Oracions subordinades

Segons la definició de la Gramàtica Essencial de la llengua catalana (IEC, 2018) la subordinació és “una relació per la qual una oració pot funcionar com ho fa un sintagma en una oració simple; l’oració subordinada depèn d’una altra oració, la principal, o d’un dels seus constituents.” Per defecte, les oracions subordinades es segmenten en ADUs independents. Tanmateix, hi ha criteris específics per a les oracions de relatiu, construccions condicionals i les oracions subordinades substantives en cas de les expressions epistèmiques i d’estil indirecte.

Pel que fa als marcadors de la subordinació, s’inclouen en l’ADU del membre subordinat per tal de poder recuperar el tipus de relació que s’estableix entre les ADUs.

- (8) ADU 1 | és aquesta sostenibilitat la que és important garantir
 ADU 2 | perquè la sostenibilitat no és una opció

(UJI-UPV 2019)

En aquest exemple l’ADU 2 manté una relació d’inferència respecte a l’ADU 1.

Per altra banda, les oracions subordinades de relatiu “es defineixen per la seva capacitat de modificar el nom o el sintagma nominal precedent.”(IEC, 2018); és a dir, funcionen com un adjectiu. Per aquest motiu no es segmenten per separat, perquè el contingut de la oració de relatiu és necessari per tal que l’ADU tingui significat.

- (9) ADU 1 un model en el qual la dona no té la necessitat de complir amb unes normes

(UJI-UPV 2019)

Tot i així, es pot donar el cas de que una part inclosa en l’oració subordinada de relatiu sigui el referent d’una altra ADU. En aquest cas, s’afegirà una nova ADU que contingui la oració subordinada de relatiu; aquesta nova ADU s’annotarà amb una relació de inferència respecte a l’ADU que conté l’oració completa. L’ADU que inclou únicament l’oració de relatiu serà a la que farà referència la relació de la tercera ADU.

- ADU 1 suposa un desig de les persones que fa perpetuar un rol històric de la dona
(10) ADU 2 que fa perpetuar un rol històric de la dona
ADU 3 legitima que les dones han de patir

(UMH-UAO-CEU 2019)

Les construccions condicionals expressen una relació causal amb la conjunció d'una causa hipotètica i un efecte possible. Aquestes construccions s'estructuren amb la conjunció “si”, o similars. Generalment l'efecte va després de la causa, encara que el efecte pot trobar-se anteposat a la causa en determinats casos (IEC, 2018).

A efectes de la segmentació, considerarem les construccions condicionals com dos ADUs relacionades amb una relació d'inferència.

- (11) si volem salvar el planeta
s'han d'impulsar polítiques anti natalistes 1

(UMH-UAO-CEU 2019)

Les expressions epistèmiques tenen uns criteris específics de segmentació. En primer lloc, les expressions epistèmiques són una modalitat oracional que emet el parlant per indicar la veracitat de l'afirmació que ha realitzat. Aquestes expressions venen introduïdes amb els verbs “pensar”, “creure” o “poder”, entre d'altres, i s'expressen amb una construcció subordinada.

En segon lloc, la cita en un debat serveix per introduir les paraules d'una altra persona dins del discurs. Aquest recurs es pot fer servir per donar validesa als arguments citant a una persona docta en l'àmbit del debat o per donar un altre punt de vista citant una persona d'un altre àmbit. Les cites poden reproduir textualment les paraules de la persona citada, estil directe, o bé parafrasejar el contingut de les paraules de la persona citada (estil indirecte).

Per la segmentació d'aquestes oracions, tant epistèmiques com cites indirectes, tindrem en compte un criteri: l'emissor de la informació de l'oració. Si l'emissor de la informació és una font externa, mantindrem tota l'oració en una ADU (exemple 12). En canvi, si l'emissor de la informació és el propi parlant, només es mantindrà la part de l'oració subordinada (exemple 13).

(12) el neonatòleg nils bergman diu que el pitjor que li pot passar a un nounat és que li separen de la seua mare (URL-UMH 2019)

(13) (jo crec que)⁵ hi ha bastants col·lectius al món que estem lluitant (UB-UOC 2019)

2.4.2.3. Exemples i metàfores

Durant l'acte de l'argumentació pot ser útil deixar de banda els arguments abstractes i les idees més generals i centrar un argument en fets materials que els receptors de l'enunciat puguin relacionar amb el context en el món real. Perelman i Olbrechts-Tyteca (1989) anomenen “enllaços que fonamenten l'estructura real” a aquest tipus d'arguments i en distingeixen tres casos: el cas particular (exemple), l'analogia i la metàfora.

En el cas particular podem trobar exemples de situacions reals, models e il·lustracions de situacions del context dels debatents. Aquestes situacions es poden utilitzar per extreure generalitzacions aplicables a un argument, normalment posterior, o per fer empatitzar a l'interlocutor amb la situació narrada en el cas particular.

(14) gamyi és una noia nascuda per gestació subrogada // a tailàndia tailàndia és un país que no té // cap tipus de regulació sobre la gestació subrogada // què va passar amb gamyi gamyi tenia una germana // varen néixer bessons // molt bé gamyi va néixer amb síndrome de down // va néixer amb una discapacitat // els pares d'austràlia van anar // a tailàndia i quan van veure que la gamyi tenia síndrome // de down el que van fer és emportar-se a austràlia // el nadó que havia nascut sense discapacitats i la gamyi va quedar a tailàndia // amb la seva mare subrogada (UB-UOC 2019)

En aquest exemple, l'equip en contra de la gestació subrogada narra la història de la nena tailandesa Gamyi i de la seva mare, amb l'objectiu de que els receptors sentin empatia cap aquest cas i el generalitzin a tota la gestació subrogada.

Per altra banda, tenim l'analogia, que és un recurs retòric que ens permet extreure una semblança entre dos fets a partir d'una lògica inductiva.

⁵No es segmenta.

- (15) per tant segons aquesta afirmació el primer negre que a estats // units se va saltar la llei de què no es podia seure en // el lloc on s'ha d'asseure una persona blanca // també ho va fer malament (UB-UOC 2019)

Amb aquest enunciat l'equip a favor va respondre a l'argument de l'equip contrari, segons el qual la gestació subrogada és il·legal i, per tant, saltar-se aquesta llei és un fet delictiu i dolent. El que volien explicar amb aquest argument és que en aquests dos casos, malgrat ser ill·legals, es lícit infringir la llei perquè es consideren una injustícia.

L'últim recurs per estructurar la realitat és la metàfora, que trasllada un significat d'una locució a una altra, com en l'exemple 16.

- (16) com l'arbre la nostra configuració es va modelant durant tota la vida adquirint idees valors i actituds socials (UdL-UIB 2019)

Tots els casos d'estructuració de la realitat, exemple, analogia o metàfora, es segmentaran com una sola ADU. Però si una ADU posterior es relaciona amb una part concreta de l'ADU segmentada, en aquest cas es crearà una nova ADU amb el segment referenciat.

2.4.2.4. Altres criteris

Hi ha altres criteris que s'han de tenir en compte per a la segmentació de les ADUs: diferencia entre connectors i marcadors, interrogacions i aparició de verbs desideratius.

Pel que fa als connectors i marcadors, hem establert el criteri d'incloure en l'ADU els connectors que relacionen les oracions i paràgrafs, però no els marcadors textuais. Per distingir els connectors i els marcadors s'utilitzaran els criteris del document de criteris lingüístics de la Universitat Politècnica de Catalunya.⁶

- (17) a. primerament [castigaríem amb el pes de la llei] (UAB-UA 2019)
b. [per tant el principal mètode substitutiu cau] (UdG-UIB 2019)

⁶<https://www.upc.edu/slt/ca/recursos-redaccio/criteris-linguistics/frases-lexic-paragraf/marcadors-i-connectors>

Respecte a les interrogacions, distingim entre les preguntes en general i les preguntes confirmatòries (*tag questions*). Les preguntes de caràcter general, no a una ADU concreta, en principi es segmenten perquè poden servir de suport d'una ADU (exemple). Les preguntes confirmatòries no es consideren (del tipus “si?” o “no?”) no es consideren una interrogació pròpiament dita: per tant, només es té en compte el text anterior.

- (18) a. és legítim poder comprar-ho tot (UMH-UAO-CEU 2019)
- b. no creieu que és molt perillós legalitzar totes aquelles realitats que existeixen (UJI-UPV 2019)

- (19) [quan mireu aquí què veieu cinquanta euros] no (UJI-UPV 2019)

Les construccions amb verbs desideratius, del tipus “vull”, “m’agradaria” o “desitjo”, no es consideren ADUs i, per tant, no es segmenten.

- (20) m’agradaria vore les proves que demostren que ha sigut denegada (URL-UMH 2019)

2.5. Anotació

En aquesta secció es descriuran les etiquetes utilitzades per a l'anotació del corpus. Les etiquetes utilitzades tracten de reproduir les característiques de l'anotació amb OVA.

2.5.1. Identificador de l'ADU

L'identificador de l'ADU és un número assignat a cada ADU de manera única i inequívoca, és a dir, cada ADU té el seu número i no es repeteix en cap altra ADU.

L'identificador s'assigna de manera automàtica en el full de càlcul i serà cronològic i correlatiu, de manera que no pot haver-hi salts en la numeració dels identificadors. Per tant, la primera ADU d'un debat tindrà l'identificador 1, la següent ADU l'identificador 2 i així successivament.

2.5.2. Part del debat

En la part del debat s'anota a quin dels quatre torns del debat pertany l'ADU. Les etiquetes utilitzades són: INTRO, ARG1, ARG2 i CONCLUS. Aquestes etiquetes corresponen a les següents parts del debat:

1. Introducció: Correspon a la primera fase dels debats, on els debatents presenten les idees principals en vers les quals estarà centrada la seva argumentació. S'etiqueta com *INTRO*.
2. Refutació: És la part central del debat on els debatents desenvoluparan els arguments exposats en la introducció i intentaran refutar els arguments de l'equip contrari. Aquesta és l'única part del debat on l'equip contrari pot intervenir en el torn de refutació de l'equip que està argumentant per qüestionar el seu argument. S'etiqueta com *ARG1* o *ARG2*.
3. Conclusions: És l'última part del debat on s'exposen les últimes idees. S'etiqueta com *CONCLUS*.

2.5.3. Número d'argument

A l'hora de plantejar un debat es fan servir unes idees centrals en les quals es centrarà l'argumentació durant el debat. Generalment es presenten dues o tres idees centrals.

Aquesta etiqueta reflecteix a quina idea o argument central pertany cada ADU de cada equip. Els equips presenten aquestes idees principals durant el torn d'introducció, que no s'anota, i les desenvolupen en els arguments dels torns de refutació. A cada argument de cada equip se li assigna un número per ordre d'aparició en els torns de refutació, començant per l'1, i s'anoten les ADUs que pertanyen a aquest argument.

2.5.4. Posicionament (*stance*)

Durant la fase preliminar del debat, el cap de jutges realitza un sorteig entre els dos equips debatents en el qual es decideix tant qui comença i conclou el debat com la posició que defensarà cada equip: a favor o en contra del tema del debat.

Aquest posicionament s'anotarà en cadascuna de les ADUs per controlar a quin equip pertany. S'anotarà amb les següents etiquetes:

- FAVOUR: Si l'emissor de l'ADU és l'equip que argumenta a favor del tema que es debat.

- AGAINST: Si l'emissor de l'ADU és l'equip que argumenta en contra del tema que es debat.

2.5.5. Identificador de la relació

Aquesta etiqueta està relacionada amb l'identificador de l'ADU (apartat 2.5.5) perquè s'utilitza per indicar amb quines ADUs té relació l'ADU que s'està anotant. L'anotació consisteix a assenyalar els números identificadors de les ADUs amb las que es relaciona l'ADU que se està anotant. Si són diverses ADUs, s'anotaran separant els seus identificadors amb el signe “;”.

2.5.6. Tipus de relació

L'última etiqueta de l'anotació, possiblement la més important de tota la tasca, consisteix a anotar el tipus de relacions que estableix l'ADU amb altres ADUs anotades prèviament. Els tipus de relacions s'anoten amb les etiquetes RA, MA, CA.

L'etiqueta RA indica una relació que, per inferència lògica, uneix una o més ADUs. Aquesta és la categoria de tipus de relació més àmplia i en ella s'inclouen relacions entre ADUs que modifiquen el significat de l'argumentació de diferents maneres. Com a norma general considerem que les oracions subordinades, causals, consecutives i condicionals estableixen una relació d'inferència respecte l'oració principal.

- | | | | | |
|------|-------|--|---|----|
| | ADU 1 | un model de barra lliure | | |
| (21) | ADU 2 | un model on tot sigui a costa de doblers | 1 | RA |
| | ADU 3 | un sistema regulat i amb controls | 1 | RA |
- (UdG-UIB 2019)

L'etiqueta MA defineix el tipus de relació que estableix una ADU quan reformula el contingut d'una altra. Per tant, aquesta ADU té un contingut similar, però no el mateix. La reformulació expressada per l'ADU pot ser de tipus generalitzador o especificador.

- | | | | | |
|------|-------|--|---|----|
| | ADU 1 | entenem com a maternitat l'acció d'exercir de mare | | |
| (22) | ADU 2 | és a dir l'acompanyament dels fills des del naixement pel seu correcte desenvolupament | 1 | MA |
- (UB-UOC 2019)

(23) 1 parleu de la regulació de la gestació subrogada
2 però és que ja està regulada 1 CA

	ADU 1	la dona és lliure d'escollir		
(24)	ADU 2	en el moment en què l'estat no li permet fer-ho l'està objectivitzant	1	RA
	ADU 3	no li deixen decidir sobre una informació que estarà conscienciada	1;2	RA

23

Un cop identificats aquests problemes, es van establir els criteris generals de mantenir la varietat del debatent i revisar de forma més exhaustiva el text contingut en les ADUs. En relació a la qualitat de l'àudio, es va decidir que en el cas de fragments inaudibles s'intentaria transcriure les idees principals del fragment.

3.1.2. Acord en la segmentació i l'anotació

Per comprovar l'acord entre anotadors, tant com pel que fa a la definició dels segments com per l'anotació dels diferents ítems, els càlculs es van fer amb el debat UB-UOC 2019 anotat en paral·lel pels dos anotadors.

Els resultats d'aquesta prova (Taula 5) mostren una disparitat de criteris en l'ítem ADU, que correspon a la segmentació de les ADUs, com es reflecteix en els baixos nivells del valor de l'acord observat i l'índex Kappa de Cohen. Els ítems *type* i *stance* no presenten problemes perquè la seva valoració és bàsicament objectiva. Els resultats obtinguts en l'índex de la Kappa de Cohen són 0,917 i 0,979, respectivament.

Pel que fa al tret *Argumental Relation Type*, l'acord entre anotadors també és força alt (86 % d'acord observat i un índex de 0,769).

	% acord observat	Kappa de Cohen
ADU	66,3	-0,017
Arg. Relation type	86,0	0,769

Taula 5: Proves d'acord entre anotadors

Es va identificar com a causa principal de desacord en la segmentació d'ADUs el biaix d'un dels anotadors pel que fa a l'aplicació dels criteris de segmentació. Per tant, es van revisar els criteris de segmentació, es va reanotar l'arxiu i es va repetir el càlcul entre anotadors (Taula 6).

	% acord observat	Kappa de Cohen
ADU	70,8	0,339

Taula 6: Prova d'acord després de revisar els criteris

Els resultats van millorar, però tot i així els nivells seguien sent baixos. Un dels factors que pot influir en aquests resultats és que el càlcul de l'acord entre anotadors pel que fa a la segmentació

no té en compte els desacords parcials, com per exemple que l'ADU contingui un espai de més, o un salt de línia o la presència o no d'una paraula, això fa baixar el percentatge de coincidència.

Després de posar en comú alguns casos concrets relacionats amb la segmentació, l'arxiu es va anotar per tercera vegada i la millora dels resultats va ser evident (Taula 7).

	% acord observat	Kappa de Cohen
ADU	76,6	0,626

Taula 7: Prova d'acord final

3.2. Resultats de l'anotació del corpus

Els resultats d'aquesta secció corresponen a la primera fase de l'anotació i en un futur s'ampliaran amb una segona fase. El corpus anotat conté un total de 3 385 ADUs en els 12 arxius que el componen. D'aquestes ADUs, unes 2 636 (77,87 %) contenen una relació amb altres ADUs. És a dir, que estan anotades amb una relació cap a una altra ADU o que una altra ADU està relacionada amb elles.

Per una altra banda, la identificació de les ADUs aïllades és més complexa, ja que s'han de cercar les ADUs amb la cel·la *Relation ID* buida i, a la vegada, que el seu ID no estigui a cap altre cel·la de *Relation ID* d'una altra ADU. Per dur a terme aquesta tasca s'ha programat un *script* en Bash, que es troba a l'Apèndix A, que ha donat un resultat de 749 ADUs aïllades (22,13 %).

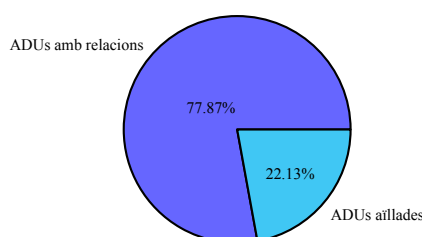


Figura 4: Recompte d'ADUs en la fase 1 del corpus Vives Debate

Aquesta diferència entre les ADUs relacionades i les aïllades és conseqüència de l'estructura de l'argumentació, un enunciat que en referencia un altre per reforçar-lo argumentalment, i també perquè, en ser debats competitius, reglats i amb puntuació, els debatents intenten desenvolupar l'argumentació per obtenir la major puntuació possible. Aquest últim fet es contraposa amb

allò que s'observa en els debats polítics, on els arguments no es desenvolupen tant, sinó que intenten transmetre una idea clara i concisa.

Per altra banda, s'ha fet un recompte dels tipus de relacions existents en el corpus, etiquetats amb RA, MA, CA. Els resultats obtinguts d'aquesta cerca han estat 2 259 inferències (RA), 300 reformulacions (MA) i 281 conflictes (CA).

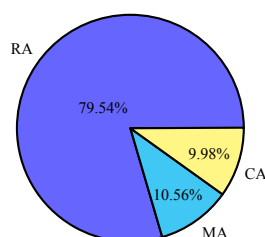


Figura 5: Recompte dels tipus de relacions en la fase 1 del corpus Vives Debate

Com es pot veure en la Figura 5, el tipus de relació més freqüent és la inferència (RA). La reformulació (MA) és menys freqüent que la inferència perquè els debatents, en comptar amb un temps limitat, intenten reformular el menys possible per tenir més temps per estructurar els arguments fent inferències. El tipus de relació menys freqüent és el conflicte, perquè el seu ús es troba bastant limitat, generalment acotat en els torns de refutació i puntualment en les refutacions. A més, les reformulacions i conflictes es desenvolupen en altres ADUs, això fa que una ADU que expressa una relació de reformulació o conflicte es desenvolupi en altres ADUs que es relacionen amb l'ADU per inferència, creant grups i cadenes d'ADUs en torn a l'ADU que es reformula o amb la que s'entra en conflicte.

A la Taula 8 oferim un resum de les ADUs anotades i les relacions establertes entre elles.

	ADUs					Relacions					
	Num. ADUs	Num. ADUs aïllades	ADUs relacionades			RA		MA		CA	
Debat UPF-UJI 2019	309	86	27,83 %	223	72,17 %	207	77,24 %	21	7,84 %	40	14,93 %
Debat URL-UMH 2019	268	102	38,06 %	166	61,94 %	173	73,31 %	30	12,71 %	33	13,98 %
Debat UAB-UdG 2019	237	138	58,23 %	99	41,77 %	190	88,79 %	17	7,94 %	7	3,27 %
Debat UAB-UA 2019	306	21	6,86 %	285	93,14 %	214	81,99 %	38	14,56 %	9	3,45 %
Debat UB-UdL 2019	196	17	8,67 %	179	91,33 %	151	83,43 %	23	12,71 %	7	3,87 %
Debat UdG-UIB 2019	218	23	10,55 %	195	89,45 %	154	87,50 %	16	9,09 %	6	3,41 %
Debat UJI-UPV 2019	297	2	0,67 %	295	99,33 %	229	88,08 %	19	7,31 %	12	4,62 %
Debat UMH-UAO-CEU 2019	277	3	1,08 %	274	98,92 %	192	80,33 %	34	14,23 %	13	5,44 %
Debat UPF-UPV 2019	350	126	36,00 %	224	64,00 %	224	79,43 %	19	6,74 %	39	13,83 %
Debat UV-UJI 2019	374	108	28,88 %	266	71,12 %	227	73,46 %	31	10,03 %	51	16,50 %
Debat UB-UOC 2019	322	60	18,63 %	262	81,37 %	185	72,27 %	34	13,28 %	37	14,45 %
Debat UDL-UIB 2018	231	63	27,27 %	168	72,73 %	113	71,52 %	18	11,39 %	27	17,09 %
Total	3385	749	22,13 %	2636	77,87 %	2259	79,54 %	300	10,56 %	281	9,89 %

Taula 8: Resum de dades del corpus anotat

3.3. Representació en grafs

Una de les característiques del programa OVA és que permet la representació visual dels debats d'una manera senzilla, ja que la interfície genera un diagrama amb les ADUs ja segmentades. Aquest tret es va perdre en canviar el programari d'anotació per un full de càlcul.

Per recuperar aquesta funcionalitat i poder representar gràficament l'estructura que formen les relacions discursives entre arguments, s'ha investigat com generar grafs a partir del corpus anotat.

Un graf està compost per tres elements: un conjunt de nodes, la unitat fonamental del graf; un conjunt d'arestes; i relacions entre els nodes i el l'extrem d'una aresta (West, 2001). Per generar aquests grafs es fa servir el paquet de Python Networkx (Hagberg *et al.*, 2008).

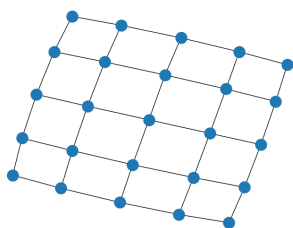


Figura 6: Exemple de graf

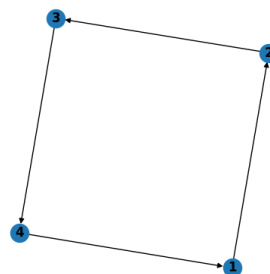


Figura 7: Exemple de digraf de 4 nodes

Els grafs generats són digrafs ordenats. Aquest tipus de grafs es caracteritzen perquè les arestes que relacionen els nodes estan dirigides. Per posar un exemple bàsic, el node 1 és 1(1,2) però el node 2 no és 2(2,1), de la mateixa manera que una ADU A estableix una relació d'inferència amb una ADU B, però la ADU B no estableix cap inferència respecte a l'ADU A. Podem veure un exemple d'aquest tipus de graf a la Figura 7

Per generar els grafs s'ha creat un *script* que selecciona diverses dades del fitxer (identificador de l'ADU, identificador de la relació i posició de l'equip) i les compila seqüencialment en el graf (Apèndix B). Cal dir que els exemples que es presenten en aquest treball només mostren la segona refutació dels equips, perquè l'elevat nombre d'ADUs dels fitxers sencers generaria més nodes, cosa que dificultaria la seva visualització.

En els grafs de les Figures 8 i 9, que corresponen a la segona refutació de l'equip a favor i dels dos equips, respectivament, es veu com els nodes formen estructures en forma de cadenes. Aquest fet reflecteix l'estructura argumental defensada pel marc teòric de Anscombe i Ducrot (apartat 1.2), que implica que un enunciat en referència un altre per reforçar-lo i que s'admeti com vàlid.

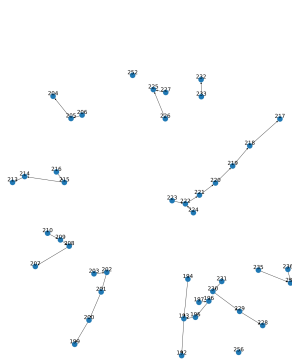


Figura 8: Refutació 2 de l'equip a favor

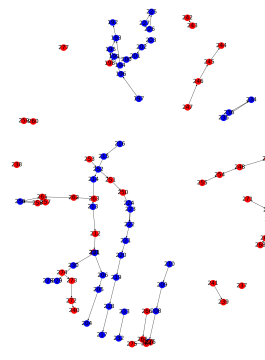


Figura 9: Refutació 2 dels dos equips

3.4. Discussió

Un dels principals problemes que trobem en l'anotació és la poca profunditat de les categories anotades, que provoca problemes de decisió i limita la informació lingüística que pot donar el corpus. Les etiquetes utilitzades defineixen categories molt àmplies pel que fa a la seva definició. Això provoca que en una mateixa categoria s'incloguin tipus diferents de relacions entre ADUs. Molt especialment, la inferència acaba sent un calaix de sastre on, per la seva

indefinició, inclouen moltes ADUs que no comparteixen característiques pragmàtiques. En el següent exemple mostrem un cas on es posa de relleu la inconcreció de les etiquetes:

- (25) ADU 1 és veritat que podem fer el que vulguem
ADU 2 però sempre respectant els límits 1
(UB-UOC 2019)

En aquest exemple podem veure com la segona ADU matisa el contingut de la primera ADU. Si intentem plasmar aquest característica, podríem arribar a anotar la segona ADU com MA. Però realment la segona ADU no és una reformulació perquè no formula la primera ADU modificant d'alguna manera el significat d'aquesta. Per tant, l'ADU hauria de ser una inferència (RA), però, com que va introduïda per una conjunció adversativa, seguint els criteris establerts s'hauria de anotar com una relació CA, encara que no acaba d'encaixar amb la noció de contradicció o conflicte en aquesta relació.

Per altra banda, el fet que l'anotació faci servir categories tan àmplies dificulta l'obtenció d'informació lingüística, però això es justifica perquè a nivell computacional interessa extreure els patrons per a l'entrenament. Els patrons s'aconsegueixen amb el major possible número de dades repetides i això s'obté amb la repetició de les temàtiques dels debats (per això es treballa amb els debats del mateix any) i amb unes categories molt generals. Especificar més les categories produiria una major divisió de les dades i caldria anotar més debats per tal que el sistema aconseguís extreure patrons.

4. CONCLUSIONS

Com s'ha vist durant tot el treball, l'anotació d'un corpus de debats implica un grau de complexitat elevat. Aquesta complexitat deriva tant de la durada dels debats com dels diversos criteris de segmentació que s'han de tenir en compte simultàniament. També afegeix dificultat el fet que els debats tenen molta interrelació: les ADUs es relacionen tant amb ADUs del mateix bloc com d'altres blocs. En conseqüència, per anotar correctament una relació s'ha de retrocedir a ADUs prèvies per veure a quina fa referència.

En segon lloc, caldria trobar una interfície que afavorís un flux de treball integrat, per evitar l'ús simultani de diverses aplicacions. Per tant, s'hauria de crear una nova eina d'anotació que

integrés el text i l'anotació o solucionar els problemes d'OVA per tal de que es poguessin fer anotacions llargues sense problemes.

Per últim, per a futures investigacions creiem que resultaria útil que l'eina d'anotació proporcionés una visualització gràfica de les anotacions. La possibilitat de veure gràficament les relacions que s'estableixen en el corpus facilita la interpretació dels patrons en les relacions. La representació d'investigacions lingüístiques amb grafs de diferents tipus es troba molt present en l'actualitat i la podem veure amb col·laboracions entre grups de lingüistes, que proporcionen el corpus i la informació lingüística, i informàtics, que dissenyen el programa necessari per representar les dades en un graf. Un exemple d'aquesta col·laboració és el recent treball Kavaz *et al.* (2021), el qual presenta models de representació per grafs del corpus de missatges tòxics per una visualització amb una interfície interactiva. Les visualitzacions tenen diferents usos, com detectar inconsistències en l'anotació o reconèixer patrons en el corpus.

BIBLIOGRAFIA

- Hagberg, A., Schult, D., i Swart, P. (2008). Exploring Network Structure, Dynamics, and Function using NetworkX. Dins de *Proceedings of the 7th Python in Science Conference (SciPy2008)*, 11-15. http://conference.scipy.org/proceedings/SciPy2008/paper_2/
- Chakrabarty, T., Hidey, C., Muresan, S., McKeown, K., i Hwang, A. (2019). AMPERSAND: Argument Mining for PERSuasive oNline Discussions. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2933-2943. <https://doi.org/10.18653/v1/D19-1291>
- Escandell Vidal, M. V. (2013). Introducció a la pragmàtica (1a ed.). Ariel.
- Esteve, M. (2020). Minería de argumentación en el Referéndum del 1 de Octubre de 2017: Argument Mining in the October 1, 2017 Referendum. *Procesamiento del lenguaje natural*. 65, 59-66.
- Han, W.-B., i Kando, N. (2019). Opinion mining with deep contextualized embeddings. Dins de NAACL -HLT 2019. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, 35-42. <http://dx.doi.org/10.18653/v1/N19-3006>
- IEC. (2018). Gramàtica essencial de la llengua catalana. <https://geiec.iec.cat/>
- IEC. (2021). Diccionari de la llengua catalana. <https://dlc.iec.cat/>
- Janier, M, Lawrence, J, i Reed, C. (2014). OVA+: An Argument Analysis Interface. *Frontiers in Artificial Intelligence and Applications*, 266, 463-464. <http://dx.doi.org/10.3233/978-1-61499-436-7-463>
- Kavaz, E., Puig, A., Rodríguez, I., Taule, M., i Nofre, M. (en premsa). Data Visualization for Supporting Linguists in the Analysis of Toxic Messages. WSCG 2021. *29th International Conference on Computer Graphics, Visualization and Computer Vision*.
- Mićović, M. (2014). *La comunicación y el discurso políticos en España y Serbia. Análisis comparativo de las estrategias argumentativas utilizadas en los debates electorales televisivos* [Tesi Doctoral, Universitat de Barcelona]. <http://hdl.handle.net/10803/285600>

- MLLP. (2020). MLLP Transcription and Translation Platform. mlp.upv.es/transcription-and-translation-platform/
- Peldszus, A., i Stede, M. (2013). From Argument Diagrams to Argumentation Mining in Texts: A Survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJ-CINI)*, 7(1), 1-31. mlp.upv.es/transcription-and-translation-platform/
- Perelman, C., i Olbrechts-Tyteca, L. (1989). *Tratado de la argumentación: La nueva retórica*. Gredos.
- Taulé, M., Antònia Martí, M., Rangel, F., i Rosso, P. (2018). Overview of the task on multi-modal stance detection in Tweets on catalan #1Oct referendum. Dins *CEUR Workshop Proceedings*, 2150, 149-166. <http://ceur-ws.org/Vol-2150/overview-Multistance18.pdf>
- Visser, J., Konat, B., Duthie, R., Koszowy, M., Budzynska, K., i Reed, C. (2020). Argumentation in te 2016 US presidential elections: Annotated corpora of television debates and social media reaction. *Language Resources and Evaluation*, 54(1), 123-154. <https://doi.org/10.1007/s10579-019-09446-8>
- Wells, G. (2018). *Minería de falacias en el discurso político*. [Treball Final de Grau]. <http://hdl.handle.net/2445/123523>
- West, D. B. (2001). *Introduction to Graph Theory* (2nd ed.). University of Illinois.
- Wikipedia. (2021). Debate. En *Wikipedia, la enciclopedia libre*. <https://es.wikipedia.org/w/index.php?title=Debate&oldid=136106295>
- Xarxa Vives. (2019). Reglament Lliga de Debat Universitària 2020. <https://www.vives.org/files/2019/09/Reglament-Lliga-de-Debat-Universit%C3%A0ria-2020.pdf>

Índex de taules

1	Dades del US2016 (Visser <i>et al.</i> , 2020)	5
2	Anys dels debats (Mićović, 2014)	6
3	Correspondència entre les sigles i les universitats dels equips debatents	8
4	Durada dels vídeos i dades lexicomètriques de les transcripcions de la primera fase d' anotació del corpus Vives Debate	9
5	Proves d' acord entre anotadors	24
6	Prova d' acord després de revisar els criteris	24
7	Prova d' acord final	25
8	Resum de dades del corpus anotat	27

Índex de figures

1	Mostra de l' anotació US2016 amb l' eina OVA	5
2	Exemple de la interfície d' anotació del programa OVA	10
3	Exemple d' un debat anotat en un full de càlcul	11
4	Recompte d' ADUs en la fase 1 del corpus Vives Debate	25
5	Recompte dels tipus de relacions en la fase 1 del corpus Vives Debate	26
6	Exemple de graf	27
7	Exemple de digraf de 4 nodes	27
8	Refutació 2 de l' equip a favor	28
9	Refutació 2 dels dos equips	28

A. *SCRIPT BASH PER A LA IDENTIFICACIÓ D'ADUS AÏLLADES*

```
1 #!/usr/bin/env bash
2
3 #Extreu les IDs de les relacions
4 awk -F',' '{print $6}' < $1 | tr ',' '\n' | tr ';' '\n' | tr -d [:alpha:]
   ↪ | tr -d [:punct:] | sort -n | uniq | tr '\n' '|' > $1.rel
5
6 #Extreu les IDs amb el camp d'IDs relacionades buit
7 awk -F',' '{if (length($7)==0 && ($2~"INTRO" || $2~"ARG1" || $2~"ARG2" ||
   ↪ $2~"intro" || $2~"arg1" || $2~"arg2")) print $1}' < $1 | tr -d
   ↪ [:alpha:] | tr -d [:punct:] > $1.id
8
9 # Extreu les IDs que no tenen correlat a les relacions
10 egrep -vw -f $1.rel < $1.id > $1.aïllat
```

B. CODI DE L'SCRIPT PER GENERAR GRAFS

```
1 import csv
2 import networkx as net
3 import matplotlib.pyplot as plt
4 from pathlib import Path
5 G = net.OrderedDiGraph()
6 GP = net.OrderedDiGraph()
7 color_map = []
8 ID = []
9 data_folder = Path("ruta_del_fitxer")
10 def id (path):
11     """Extreu les dades de la columna ID i les guarda a la llista
12     ↪ ID.
13     path -- ruta del fitxer csv.
14     """
15     with open(path, 'r') as csv_file:
16         csv_reader = csv.reader(csv_file, delimiter=',')
17         for row in csv_reader:
18             ID.append(row[0])
19             pass
20 def nodes(graph, path):
21     """Aplica la funció id i guarda les IDs i els colors dels nodes,
22     ↪ en funció de la columna 3, en el graf.
23     graph -- El graph on es guarden les dades.
24     path -- ruta del fitxer.
25     """
26     ID.clear()
27     id(path)
28     with open(path, 'r') as csv_file:
29         csv_reader = csv.reader(csv_file, delimiter=',')
30         for row in csv_reader:
```

```

29         if row[3]=='FAVOUR':
30             tmp = row[0]
31             graph.add_node(int(tmp))
32             color_map.append("blue")
33         elif row[3] == 'AGAINST':
34             tmp = row[0]
35             graph.add_node(int(tmp))
36             color_map.append("red")
37
38     def edges(graph, path):
39         """Extreu els IDs relacionants de la columna 5 i els guarda en
40         ↪ el graf.
41         graph -- El graph on es guarden les dades.
42         path -- ruta del fitxer.
43         """
44         with open(path, 'r') as csv_file:
45             csv_reader = csv.reader(csv_file, delimiter=',')
46             line_count = 0
47             for row in csv_reader:
48                 if row[5] == '':
49                     pass
50                 elif row[5] not in ID:
51                     pass
52                 else:
53                     graph.add_edge(int(row[0]),int(row[5]))
54
55     # # Genera un graf d'un equip
56     prova1 = data_folder / "fitxer1.csv"
57     nodes(G, prova1)
58     edges(G, prova1)
59     plt.figure(1, figsize=(20,20))
60     pos = net.spring_layout(G, k=0.2)
61     net.draw_networkx(G, pos, font_size=18, verticalalignment= 'bottom')

```

```
60 plt.savefig('ruta/fitxer1.svg')
61
62 ### Genera un graf dels dos equips
63 color_map.clear()
64 prova2 = data_folder / "fitxer2.csv"
65 nodes(GP, prova2)
66 edges(GP, prova2)
67 plt.figure(2, figsize=(20,20))
68 pos2 = net.spring_layout(GP, k=0.5, iterations=200)
69 net.draw(GP, pos2, node_color=color_map, node_size=500, with_labels=True,
    ↪ font_size=20)
70 plt.savefig('ruta/fitxer2.svg')
```
