



Discourse markers and other signals: annotation and analysis

Ludivine CRIBLE

Bucharest, 15-16 Oct 2019

A stylized green feather graphic is positioned on the left side of the slide, extending from the bottom towards the top. It has a central rachis with numerous barbs branching out, creating a fan-like shape. The color is a muted green.

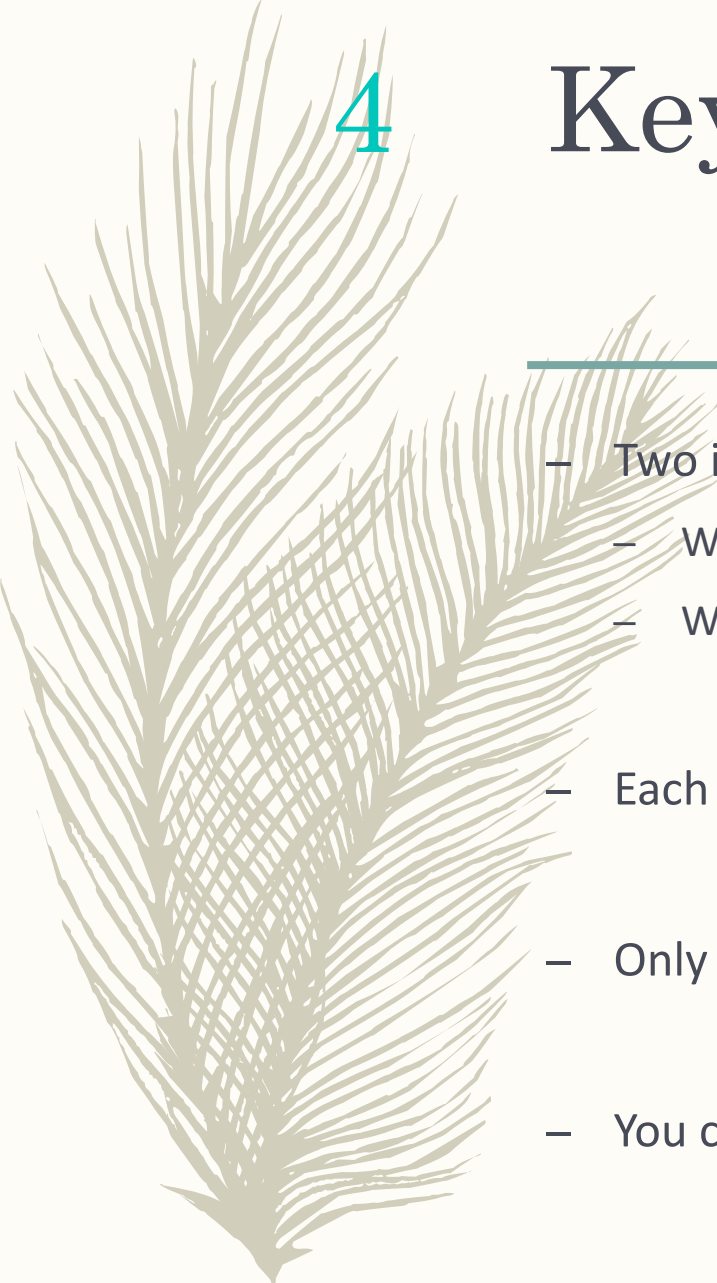
2

Overview

1. Domains and functions : operational definitions
2. EXMARaLDA suite: general functionalities
3. Hands-on demo : creating an annotated TedTalk transcription
4. Extracting and analyzing data
5. The next step : signalling analysis

The taxonomy in practice

Definitions



4

Key principles : reminders

- Two independent layers of functional information
 - What is the relation/function expressed by the (semantics of the) DM ? ➔ 15
 - Which type of content/elements/layer does the DM target ? ➔ 4
- Each function can combine with each domain (theoretically)
- Only 1 value per level (no double tags)
- You can start annotating at any level

5

Domains

Ideational	Rhetorical	Sequential	Interpersonal
Objective relations between external facts Low degree of speaker involvement Incompatible with expressions of opinion	Subjective relations between thoughts or speech-acts Speaker's attitude, beliefs, reasoning Distance from facts ("I think that...", "I can say that...")	Segments management Structuring topics, turns, digressions, hesitations, stalling Make the steps and flow of speech more explicit	Addressee management Phatic function, manifests the relationship with the hearer Explicit call or answer to the addressee

6

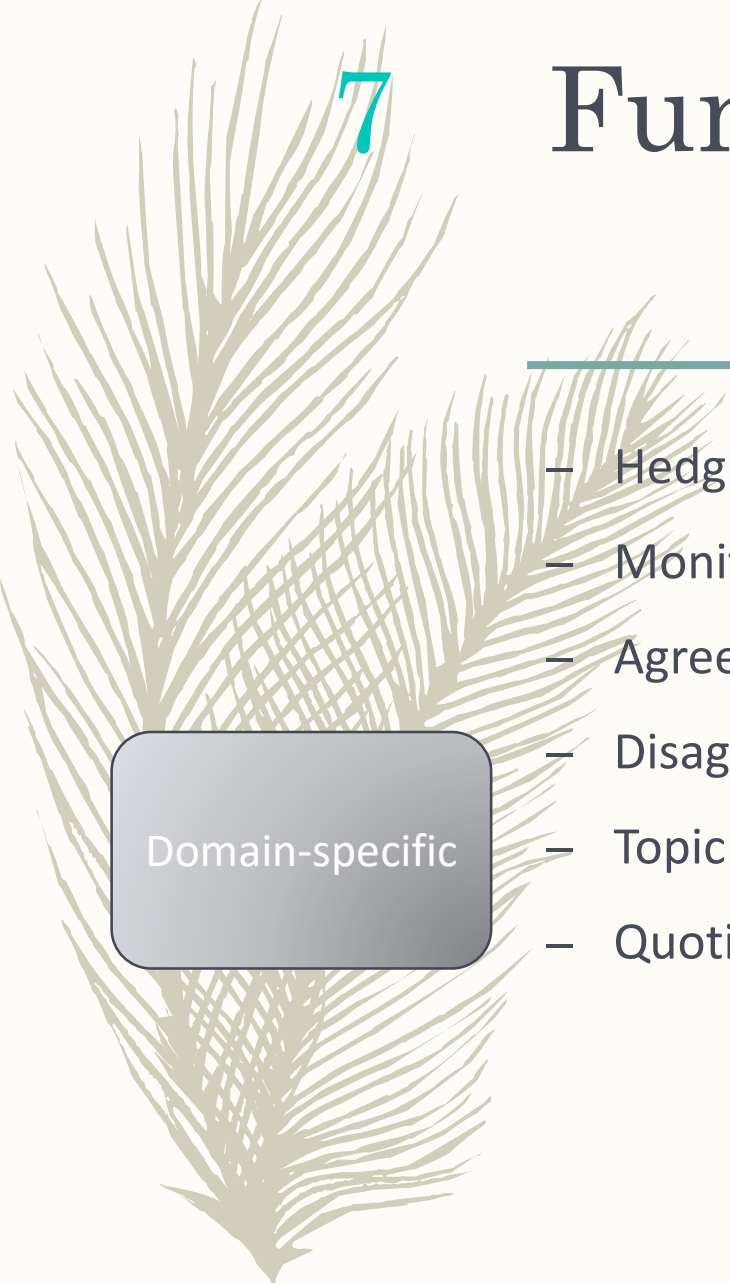
Functions I : discourse relations

Conjunction

Contingency

Comparison

- Addition (ADD) : S2 provides discourse-new information related to S1
- Specification (SPE) : S2 elaborates on S1 with more details or an example
- Temporal (TMP) : the two segments are chronologically ordered
- Cause (CAU) : S2 explains the situation in S1
- Consequence (CSQ) : S2 is the result of the situation in S1
- Condition (CND) : S2 is the condition for the truth/relevance of S1
- Concession (CCS) : S2 denies expectations related to S1
- Contrast (CTR) : the two segments differ w.r.t a shared property
- Alternative (ALT) : the segments can replace each other



7

Functions II : speech-specific

- Hedging (HDG) : the DM signals some approximation
- Monitoring (MNT) : the DM signals the speaker's intent to control the flow
- Agreeing (AGR) : the DM signals agreement
- Disagreeing (DIS) : the DM signals disagreement
- Topic (TOP) : the DM signals a start, change or return to topic
- Quoting (QUO) : the DM introduces (pseudo-)reported speech

Domain-specific

8

Examples

	IDE	RHE	SEQ	INT
Addition	le grand frère avait un rôle de papa et en plus d'être papa il avait un rôle de d'essayer les choses avant nous	non je marchais pas ah non non j'ai pas couru (0.180) et j'ai fait encore un détour	Pacs avait fait une intendance aux baladins (0.780) et euh Camille lui dit euh tu oublieras pas de payer	<spk1> tu dis euh cheese pour le cliché et genre euh un peu pour se cacher <spk2> et un peu pour se cacher aussi ouai
Alternative	on est plusieurs ou tu me vouvoies ?	c'est pas pour ça qu'on fait de la musique mais c'est enfin c'est pas pour être reconnu dans la rue	euh ben j'ai fait euh deux ans enfin ma première et ma deuxième euh d'institutrice euh primaire	<spk1> j'avais repris euh des études en gestion des ressources humaines [...] <spk2> directement après? <spk1> ben euh enfin j'ai arrêté euh l'année passée euh avril et euh [...] l'année scolaire suivante
Concession	elle devait partir le lendemain mais elle n'est jamais partie	si la démocratie est un mot ancien, ici et maintenant la démocratie signifie la prospérité pour tous	c'était assez comique de les entendre parler comme ça euh des filles (0.690) mais euh ouais puis après euh voilà quoi	cet auditeur euh vigilant il va vous dire tiens euh encore Jean d'Ormesson mais on entend Jean d'Ormesson à chaque automne



9

Tips and notes

- Domains form a **relative** cline, allow for “more” or “less” interpretations
- Domains might not mean exactly the same thing for all functions, be **flexible**
- In case of doubt for the function, the **bias** is the “dictionary” meaning
- Test phase and discussion with second annotator necessary
- Practice makes perfect 😊

EXMARaLDA suite

General functionalities

Generalities

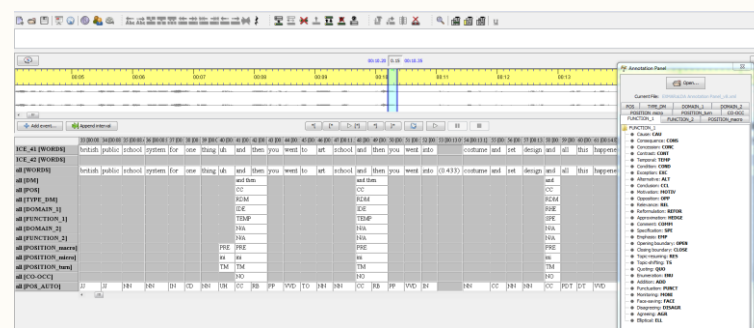
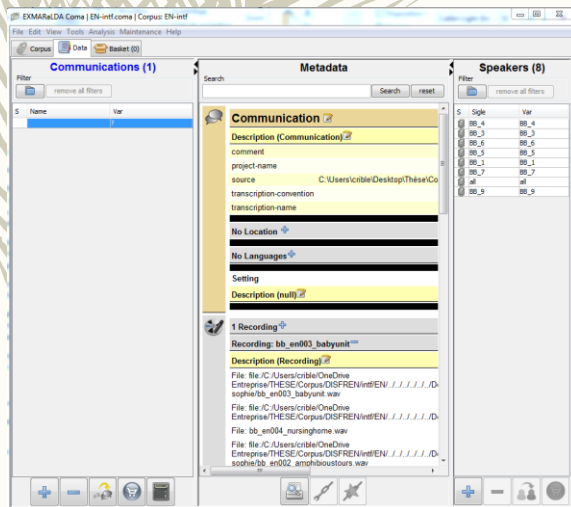
- Thomas Schmidt's team in Hamburg (CLARIN-D)
- Open-source annotation software
- Designed specifically for spoken text
 - transcription
 - text-to-sound alignment
 - annotation
- Download and documentation available at: <http://exmaralda.org/en/>

1

2

(3)

- Corpus Manager for corpus metadata
- Partitur Editor for transcription and annotation
- Exakt for extraction/concordancer

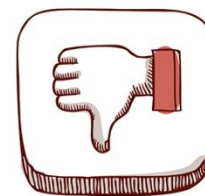


Left Context	Match	Right Context	DM	POS	TYPE_DM	DOANH_1	FUNCTION_1	DOANH_2	FUNCTION_2	Filename[1]
un sentiment et il nous munit par avance de polices.	après que	lors d'un complot manifesté de son fait quand même.	après que	CC	RDIM	IDM	CAU	N/A	N/A	FFB-entr-01
(0.84) et que (0.64) tout en se déplaçant vers l'ouest.	après que	il a d'abord gardé la même position et c'est	après que	CC	RDIM	IDM	CAU	N/A	N/A	FFB-entr-01
ont (13.64) tout nous munit par avance (0.172) alors	après que	lors d'un complot manifesté (0.87) qui est de plus	après que	RB	RDIM	IDE	ENU	N/A	N/A	FFB-entr-01
déjà la probabilité (0.676) et qui est de plus en plus	après que	manifestation (0.42) nous que ça passait (0.397) et	après que	CC	RDIM	IDE	MOTIV	N/A	N/A	FFB-entr-01
travaux (0.47) et que (0.47) nous munit par avance	après que	CC -> nous munit par avance (0.47) nous munit	après que	CC	RDIM	IDE	MOTIV	N/A	N/A	FFB-entr-01
également des étonnement souffrant (12.042) et	après que	il a quand même accablé (0.033) à l'insuffisance	après que	CC	RDIM	SEQ	ADJO	N/A	N/A	FFB-entr-01
trouva à nos (0.88) et les de l'armée (13.101) e	après que	il a assumé son trépas par quand qu'on ne s'est	après que	VP	RDIM	RPE	SPE	N/A	N/A	FFB-entr-01
0.103) et que (0.47) nous munit par avance	après que	il a été à cette occasion de l'absence (0.703) et	après que	CC	RDIM	MOTIV	CAU	N/A	N/A	FFB-entr-01
j'en (0.87) et de d'assurer au sein trépas par que	après que	il a été à cette barière de sensibilité (0.763) que tradit	après que	CC	RDIM	IDE	TEMP	N/A	N/A	FFB-entr-01
(0.428) et c'est comme l'ambulant c'est par nos	après que	(12.574) par que le genre (0.282) qui est	après que	UH	NFDM	INT	MONI	N/A	N/A	FFB-entr-01
est et est et est et est et est et est et est et	après que	est et est et est et est et est et est et est et	après que	CC	RDIM	IDE	CAU	N/A	N/A	FFB-entr-01
est qu'il a quand même (0.712) déconvoqua le	après que	est (0.4) à (0.408) à déplié (1) l'engagement	après que	CC	RDIM	RPE	SPE	N/A	N/A	FFB-entr-01
est (0.4) à (0.408) à déplié (1) l'engagement	après que	est et est et est et est et est et est et est et	après que	CC	RDIM	SEQ	ADJO	N/A	N/A	FFB-entr-01
est (0.408) à déplié (1) l'engagement	après que	est et est et est et est et est et est et est et	après que	CC	RDIM	SEQ	ADJO	N/A	N/A	FFB-entr-01
est (0.408) à déplié (1) l'engagement	après que	est et est et est et est et est et est et est et	après que	CC	RDIM	SEQ	ADJO	N/A	N/A	FFB-entr-01
(0.607) est et que donc c'est toujours trop évident	après que	il a dit une chose sur l'ouest qui est tellement	après que	CC	RDIM	SEQ	ADJO	N/A	N/A	FFB-entr-01
la à la j'en trouve la de l'aire de Cléaire et c'est	après que	le genre abstrait (0.813) se serait - ce qui (0.103) une	après que	VP	NFDM	RPE	PUNCT	N/A	N/A	FFB-entr-01
est (0.4) à (0.408) à déplié (1) l'engagement	après que	est et est et est et est et est et est et est et	après que	CC	RDIM	SEQ	ADJO	N/A	N/A	FFB-entr-01
par si l'est entre les mille - nous - sont - trépas	après que	il a dit une chose sur l'ouest qui est tellement	après que	CC	RDIM	SEQ	ADJO	N/A	N/A	FFB-entr-01
est (0.4) à (0.408) à déplié (1) l'engagement	après que	est et est et est et est et est et est et est et	après que	CC	RDIM	SEQ	ADJO	N/A	N/A	FFB-entr-01
est (0.4) à (0.408) à déplié (1) l'engagement	après que	est et est et est et est et est et est et est et	après que	CC	RDIM	SEQ	ADJO	N/A	N/A	FFB-entr-01
est (0.4) à (0.408) à déplié (1) l'engagement	après que	est et est et est et est et est et est et est et	après que	CC	RDIM	SEQ	ADJO	N/A	N/A	FFB-entr-01
est (0.4) à (0.408) à déplié (1) l'engagement	après que	est et est et est et est et est et est et est et	après que	CC	RDIM	SEQ	ADJO	N/A	N/A	FFB-entr-01
est (0.4) à (0.408) à déplié (1) l'engagement	après que	est et est et est et est et est et est et est et	après que	CC	RDIM	SEQ	ADJO	N/A	N/A	FFB-entr-01
est (0.4) à (0.408) à déplié (1) l'engagement	après que	est et est et est et est et est et est et est et	après que	CC	RDIM	SEQ	ADJO	N/A	N/A	FFB-entr-01
est (0.4) à (0.408) à déplié (1) l'engagement	après que	est et est et est et est et est et est et est et	après que	CC	RDIM	SEQ	ADJO	N/A	N/A	FFB-entr-01
est (0.4) à (0.408) à déplié (1) l'engagement	après que	est et est et est et est et est et est et est et	après que	CC	RDIM	SEQ	ADJO	N/A	N/A	FFB-entr-01
est (0.4) à (0.408) à déplié (1) l'engagement	après que	est et est et est et est et est et est et est et	après que	CC	RDIM	SEQ	ADJO	N/A	N/A	FFB-entr-01
est (0.4) à (0.408) à déplié (1) l'engagement	après que	est et est et est et est et est et est et est et	après que	CC	RDIM	SEQ	ADJO	N/A	N/A	FFB-entr-01
est (0.4) à (0.408) à déplié (1) l'engagement	après que	est et est et est et est et est et est et est et	après que	CC	RDIM	SEQ	ADJO	N/A	N/A	FFB-entr-01
est (0.4) à (0.408) à déplié (1) l'engagement	après que	est et est et est et est et est et est et est et	après que	CC	RDIM	SEQ	ADJO	N/A	N/A	FFB-entr-01
est (0.4) à (0.408) à déplié (1) l'engagement	après que	est et est et est et est et est et est et est et	après que	CC	RDIM	SEQ	ADJO	N/A	N/A	FFB-entr-01
est (0.4) à (0.408) à déplié (1) l'engagement	après que	est et est et est et est et est et est et est et	après que	CC	RDIM	SEQ	ADJO	N/A	N/A	FFB-entr-01
est (0.4) à (0.408) à déplié (1) l'engagement	après que	est et est et est et est et est et est et est et	après que	CC	RDIM	SEQ	ADJO	N/A	N/A	FFB-entr-01
est (0.4) à (0.408) à déplié (1) l'engagement	après que	est et est et est et est et est et est et est et	après que	CC	RDIM	SEQ	ADJO	N/A	N/A	FFB-entr-01
est (0.4) à (0.408) à déplié (1) l'engagement	après que	est et est et est et est et est et est et est et	après que	CC	RDIM	SEQ	ADJO	N/A	N/A	FFB-entr-01
est (0.4) à (0.408) à déplié (1) l'engagement	après que	est et est et est et est et est et est et est et	après que	CC	RDIM	SEQ	ADJO	N/A	N/A	FFB-entr-01
est (0.4) à (0.408) à déplié (1) l'engagement	après que	est et est et est et est et est et est et est et	après que	CC	RDIM	SEQ	ADJO	N/A	N/A	FFB-entr-01
est (0.4) à (0.408) à déplié (1) l'engagement	après que	est et est et est et est et est et est et est et	après que	CC	RDIM	SEQ	ADJO	N/A	N/A	FFB-entr-01
est (0.4) à (0.408) à déplié (1) l'engagement	après que	est et est et est et est et est et est et est et	après que	CC	RDIM	SEQ	ADJO	N/A	N/A	FFB-entr-01
est (0.4) à (0.408) à déplié (1) l'engagement	après que	est et est et est et est et est et est et est et	après que	CC	RDIM	SEQ	ADJO	N/A	N/A	FFB-entr-01
est (0.4) à (0.408) à déplié (1) l'engagement	après que	est et est et est et est et est et est et est et	après que	CC	RDIM	SEQ	ADJO	N/A	N/A	FFB-entr-01
est (0.4) à (0.408) à déplié (1) l'engagement	après que	est et est et est et est et est et est et est et	après que	CC	RDIM	SEQ	ADJO	N/A	N/A	FFB-entr-01
est (0.4) à (0.408) à déplié (1) l'engagement	après que	est et est et est et est et est et est et est et	après que	CC	RDIM	SEQ	ADJO	N/A	N/A	FFB-entr-01
est (0.4) à (0.408) à déplié (1) l'engagement	après que	est et est et est et est et est et est et est et	après que	CC	RDIM	SEQ	ADJO	N/A	N/A	FFB-entr-01
est (0.4) à (0.408) à déplié (1) l'engagement	après que	est et est et est et est et est et est et est et	après que	CC	RDIM	SEQ	ADJO	N/A	N/A	FFB-entr-01
est (0.4) à (0.408) à déplié (1) l'engagement	après que	est et est et est et est et est et est et est et	après que	CC	RDIM	SEQ	ADJO	N/A	N/A	FFB-entr-01
est (0.4) à (0.408) à déplié (1) l'engagement	après que	est et est et est et est et est et est et est et	après que	CC	RDIM	SEQ	ADJO	N/A	N/A	FFB-entr-01
est (0.4) à (0.408) à déplié (1) l'engagement	après que	est et								

13

Pros and cons

- Open-source
- All-in-one
- User-friendly, intuitive (vs. Praat)
- Few constraints (vs. ELAN)
- Interoperable format
- Cannot handle heavy files
- Several steps for extraction
- Each annotation tier per speaker



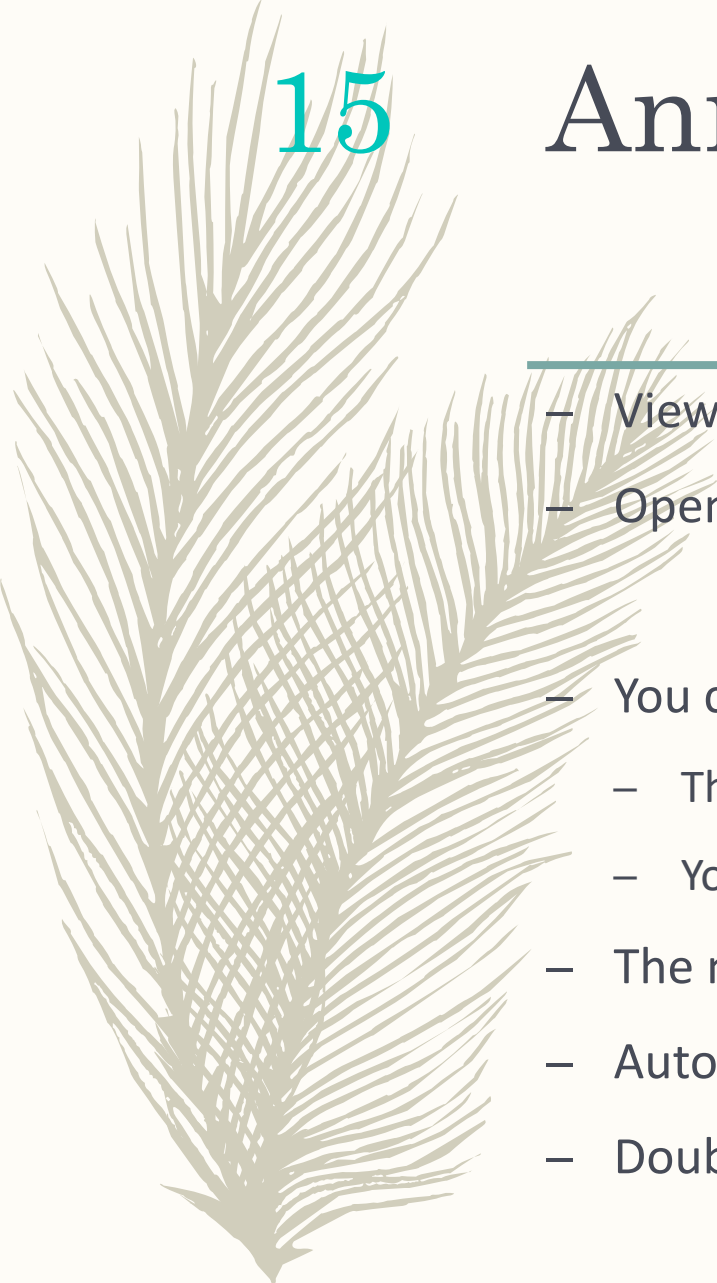


14

Input formats

- ELAN (.eaf)
- Praat (.TextGrid)
- Transcriber (.trs)
- Folker (.flk)
- CHAT (.cha)
- Anvil (.anvil)
- Annotation Graph file (.xml)
- Plain text (.txt)
- Treetagger (.txt)
- TEI (.xml)

same formats available
for export



15

Annotation panel

- View > Annotation panel
- Open : choose your `.xml` file in its folder
- You can edit the annotation panel with any text editor, e.g. Notepad++
 - The file provided follows Crible & Degand (in press)
 - You can change it or create a new one → cf. EXMARaLDA documentation
- The name of the « category » must be exactly the same as the name of the tier
- Automatically displays the list of available values + any description you want
- Double-click on the value to add it in the cell (avoids spelling mistakes)



16

Tips for DM annotation

- word-level segmentation
- either merge transcription tiers or double annotation tiers
- enter list of labels as « Annotation panel » for easy use
- prefer chronological order than DM-by-DM to understand the context
- don't do 5 hours in a row
- keep calm 😊

Creating an annotated TedTalk

Hands-on demo

Exercise 1

1. Use transcript provided or download any from <https://www.ted.com/>
2. Import it to Partitur Editor
3. Select segmentation rule
4. Create annotation tiers
5. Open annotation panel
6. Identify 5 DRDs and annotate their functions
7. Save as .exb file

Extraction and analysis

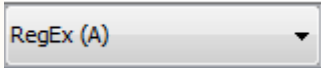
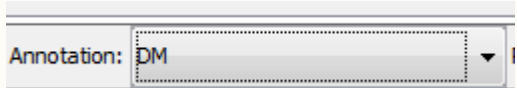
From EXMARaLDA to Excel

CorpusManager file (1)

- Group all your annotated files (.exb) in the same folder
- Open « CorpusManager » (CoMa)
- File > Create corpus from transcriptions
- Name the corpus
- Click on « Browse » : go into the folder where all the .exb files are stored
- **DO NOT CLICK ON ONE OF THE FILES**
 - click anywhere else in the folder, otherwise the corpus will erase the .exb file
 - make sure that you can read: File name > *YourCorpusName*.coma
- It will show how many transcription files you have in this folder
- « Next »

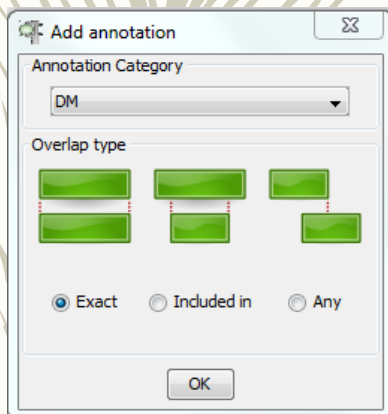
CorpusManager file (2)

- « Select transcriptions » : just click on « Next »
 - « Segmentation »
 - Tick the box on « Segment transcriptions »
 - Select « ...use default segmentation », click on « Next »
 - « Metadata assignment »
 - « Speakers »
- } Click on « Next »
- Click on « Finish » ➔ you created a .coma file

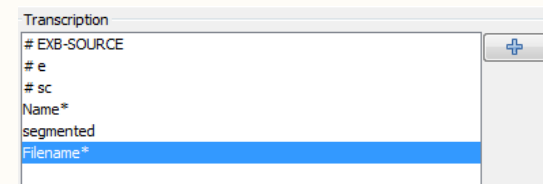
- Open EXAKT
- File > Open corpus (or shortcut) and find the .coma file you just created 
- Select « RegEx(A) » 
- Annotation: Select your tier name, e.g. « DM » 
- In the « RegEx » box, type your search string, e.g. « well »
 - Typing the dollar sign \$ will give you all the annotations, everything you typed in the « DM » tier
 - Then click on the binoculars on the right 

23

Visualizing the annotations



- You will see a concordancer with all your DMs.
- To add the annotations from other tiers:
 - Columns (top-left corner) > Add annotation
 - Select the Annotation Category you want (e.g. start with « DM », then « DOMAIN »...)
 - The « Exact » option is fine
 - « OK »
- To add metadata, such as the name of the transcription file:
 - Columns > Metadata
 - Selection « Filename* », click on the « + » sign, then « OK »



24

Visualizing the annotations

The result should look like this:

RegEx (A)		Annotation: DM		Regex: \$					
#	S	Communication	Speaker	Left Context	Match	Right Context	DM	DOMAIN_1	FUNCTION_1
1	✓		all	lo hi (0.620) could you introduce yourself please	and	tell us a little bit about what you do here (0.860	and	SEQ	ADD
2	✓		all	you do here (0.860) yes my name is Sam (0.560) um	and	I run an amphibious tour operation in Plymouth (1.	and	SEQ	ADD
3	✓		all	uth (1.170) it 's called 'ducks 'n 'drake (0.560)	right	(1.100) and (0.613) and uh yeah we run around Plym	right	INT	MONI
4	✓		all	's called 'ducks 'n 'drake (0.560) right (1.100)	and	(0.613) and uh yeah we run around Plymouth (0.910)	and	SEQ	PUNCT
5	✓		all	ducks 'n 'drake (0.560) right (1.100) and (0.613)	and	uh yeah we run around Plymouth (0.910) picking up	and	RHE	SPE
6	✓		all	n 'drake (0.560) right (1.100) and (0.613) and uh	yeah	we run around Plymouth (0.910) picking up in the c	yeah	INT	AGR
7	✓		all	ntre of town and taking people out onto the water	so	they get a (0.520) a an all encompassing view of (	so	IDE	CONS
8	✓		all	ater and on land with specialist vehicles (0.650)	right	(0.250) ok (0.660) and (0.130) you said this goes	right	INT	MONI
9	✓		all	nd with specialist vehicles (0.650) right (0.250)	ok	(0.660) and (0.130) you said this goes round in P1	ok	INT	MONI
10	✓		all	cialist vehicles (0.650) right (0.250) ok (0.660)	and	(0.130) you said this goes round in Plymouth could	and	SEQ	TS
11	✓		all	it being Europe 's largest naval base (0.380) uh	in fact	that 's why we called our (0.160) t (0.060) tour '	in fact	RHE	COMM
12	✓		all	r (0.160) t (0.060) tour 'ducks 'n 'drake (0.860)	and	what 's the significance of that Sir Francis drake	and	SEQ	OPEN
13	✓		all	of that Sir Francis drake was based here (0.750)	right	ok he led his he he was (0.110) playing bowls up o	right	INT	MONI
14	✓		all	at Sir Francis drake was based here (0.750) right	ok	he led his he he was (0.110) playing bowls up on t	ok	INT	MONI
15	✓		all	he was (0.110) playing bowls up on the ho (0.620)	when	he when he uh when he saw the uh invasion and led	when	IDE	TEMP
16	✓		all	0.110) playing bowls up on the ho (0.620) when he	when	he uh when he saw the uh invasion and led his (0.1	when	IDE	TEMP
17	✓		all	ing bowls up on the ho (0.620) when he when he uh	when	he saw the uh invasion and led his (0.170) led his	when	IDE	TEMP
18	✓		all	0) when he when he uh when he saw the uh invasion	and	led his (0.170) led his ships out to fight them (0	and	SEQ	ADD
19	✓		all	s (0.170) led his ships out to fight them (0.720)	ok	(0.560) and (0.220) what 's the importance of the	ok	INT	MONI
20	✓		all	ed his ships out to fight them (0.720) ok (0.560)	and	(0.220) what 's the importance of the 'ducks 'then	and	SEQ	TS
21	✓		all	and (0.220) what 's the importance of the 'ducks	'then	(0.150) in the name (0.490) the 'ducks 'are the sp	then	SEQ	TS
22	✓		all	'become (0.230) have become known as 'ducks ' . um	because	(x 'DUKW ' obviously sounds like 'duck 'and they '	because	IDE	CAU
23	✓		all	um because (x 'DUKW ' obviously sounds like 'duck	'and	they 're amphibious ok (0.420) h (0.020) hence the	and	RHE	SPE
24	✓		all	iously sounds like 'duck 'and they 're amphibious	ok	(0.420) h (0.020) hence the 'ducks 'in 'ducks 'n '	ok	INT	MONI
25	✓		all	20) hence the 'ducks 'in 'ducks 'n 'drake (0.950)	and	(0.440) why do you think (0.050) um this kind of t	and	SEQ	OPEN
26	✓		all	(0.360) um particularly unique or special (1.460)	well	(0.020) you 're not going to be able to get on a c	well	SEQ	OPEN
27	✓		all	le to get on a coach and then end up in the water	because	if you do you 'll sink (0.890) the uh (0.770) the	because	IDE	CAU
28	✓		all	t on a coach and then end up in the water because	if	you do you 'll sink (0.890) the uh (0.770) the (0.	if	IDE	COND
29	✓		all	) also you 're on (0.320) a bright yellow machine	so	people are gonna see you (0.520) um and and uh and	so	IDE	CONS
30	✓		all	ow machine so people are gonna see you (0.520) um	and	and uh and it 's a lot of fun (0.910) rather than	and	SEQ	ADD

25

Exploring the annotations

- You can add more characters to the Left and Right context by clicking on the magnifying glass on the right



- By double-clicking on one item, you can visualize it in the transcription format at the bottom

- You can also play it!

A screenshot of the software interface. On the left, there is a list of annotations: BB_3 [WORDS], BB_1 [WORDS], all [WORDS], and all [DM]. An arrow points from the 'all [WORDS]' annotation to a transcription window on the right. The transcription window shows a timeline with time markers [01:18:], [01:20:], [01:21:], and [01:23:]. The text in the transcription window is: and what 's the significance of that Sir Francis drake was based here (0.750) right ok. The word 'and' is highlighted in blue.

- You can add a « comment » column if you want, once you revise your annotations
 - Columns > Add analysis > Analysis name: « Comment » > OK



26

Extracting the annotations

- Click anywhere on the concordancer
- Ctrl + A (select everything)
- Ctrl + C (copy)
- Go on Excel and Ctrl + V (paste on a new Excel sheet)



27

Working under Excel

- You can now filter your data, create pivot tables and graphs, look at frequencies...

Inter- and intra-annotator reliability

- To assess the reliability and replicability of your analysis
 - Intra = repeat your annotations after a while and compare
 - Inter = compare with another annotator
- % measured in Excel : `IF((A1=A2);"same";"diff")`
- Kappa scores measured in R or online : <https://nlp-ml.io/jg/software/ira/>
- Aim for $k = 0.7$, see Spooren & Degand (2010)



Thank you for your attention !

Questions and comments welcome

Ludivine Crible

ludivine.crible@ed.ac.uk