Quantitative study of linguistic phenomenas as indices of Thought and Language Disorders

Maxime Amblard

December 8th 2014

(In)Coherence du Discours 2

LORIA - UMR 7503
Université de Lorraine

MSH-Lorraine - USR 3261
Inria Nancy Grand Est

CNRS

Motivation

- Conversational representations : involve both pragmatic and semantic representations.
- Linguistic studies of mental diseases [Chaika, 1974] and [Fromkin, 1975]
- Pragmatic discontinuities in performing verbal interaction, [Musiol and Trognon, 1996]
- Four kinds of breaking in conversations with schizophrenics : either between, or within interventions, involving two or three utterances, [Musiol and Verhaegen, 2009]

Plan

Plan

SLAM - Schizophrenics and Language : Analyse and Modelling

The project aims to systematize the study of pathological conversations under interdisciplinary approaches

- Building of a linguistic resource on mental pathology
- Epistemological and philosophical studies (norm, madness, rationality)
- Identify these purposes with :
  - formal models (as SDRT)
  - tools and methods of Natural Language Processing

SLAM - Schizophrenics and Language : Analyse and Modelling

The project aims to systematize the study of pathological conversations under interdisciplinary approaches

- Building of a linguistic resource on mental pathology
- Epistemological and philosophical studies (norm, madness, rationality)
- Identify these purposes with :
  - formal models (as SDRT)
  - tools and methods of Natural Language Processing

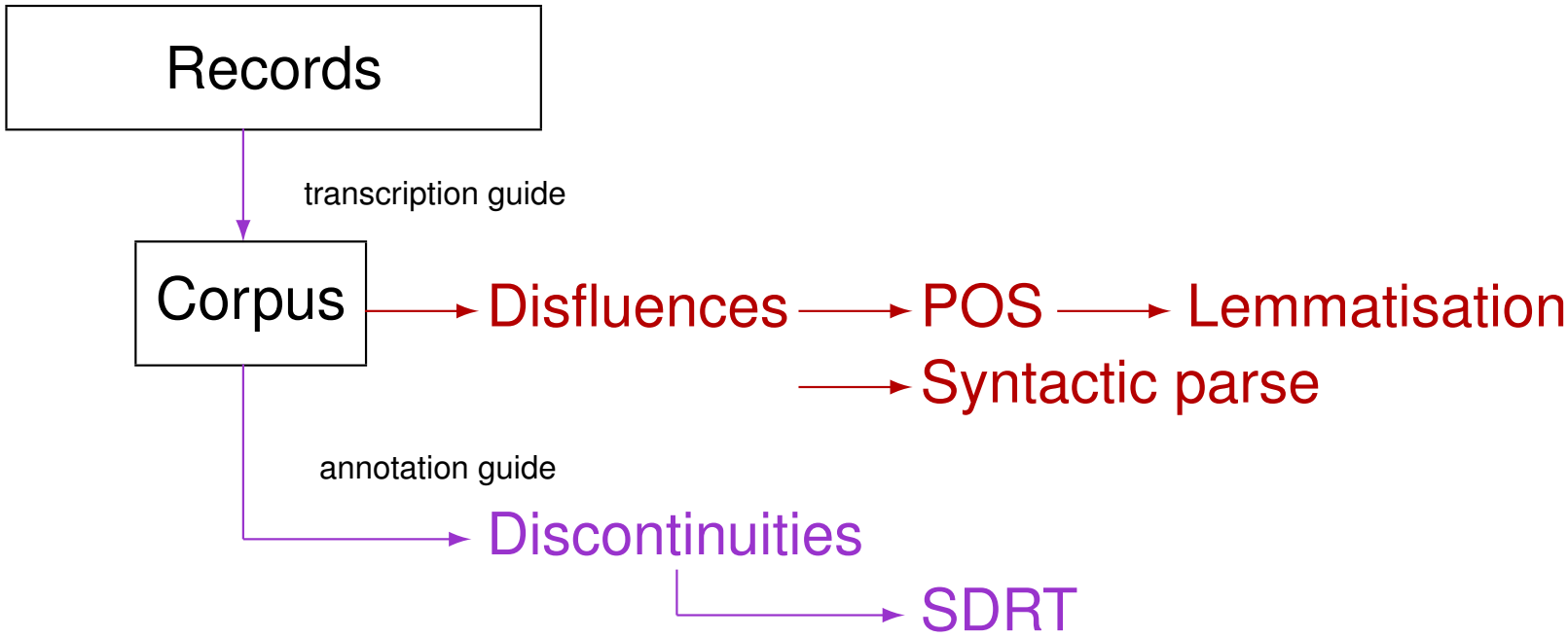
- [Rebuschi et al., 2014] : explicit correlations
- [Amblard and Fort, 2014] : pathological use of disfluencies

Some Objectives

- Define clues that can participate in the diagnosis
- Define remediation process
- Measure the impact of the remediation

- Propose a corupus (as much as possible available)
- Providing Tools for collecting the corpus

Contexte of the SLAM project



Replay linguistic ambiguities

G82 l'an dernier euh (→) j'savais pas comment faire **j'étais perdue** et pourtant j'avais pris mes médicaments j'suis dans un état vous voyez même ma bouche elle est sèche j'suis dans un triste état

I didn't know what to do. I was lost.

V83 Vous êtes quand même bien (↑)

G84 J'pense que ma tête est bien mais on croirait à moitié (↓) la moitié qui va et la moitié qui va pas j'ai l'impression de ça vous voyez (↑)

V85 D'accord

G86 Ou alors c'est la conscience peut être la conscience est ce que c'est ça (↑)

V87 Vous savez **ça arrive à tout le monde d'avoir des moments biens et des moments où on est perdu**

Everybody is lost at times.

G88 **Oui j'ai peur de perdre tout le monde**

Yes I am afraid I lose everybody.

V89 Mais ils vont plutôt bien vos enfants (↑)

G90 Ils ont l'air ils ont l'air mais ils ont des allergies ils ont (→) mon petit fils il s'est cassé le bras à l'école tout ça

Formal representation

- Two interlocutors, thus two representations of the exchange
  - psychologist : build a representation in spite of pragmatic rules
  - schizophrenic : semantically coherent, but pragmatically incorrect

Formal representation

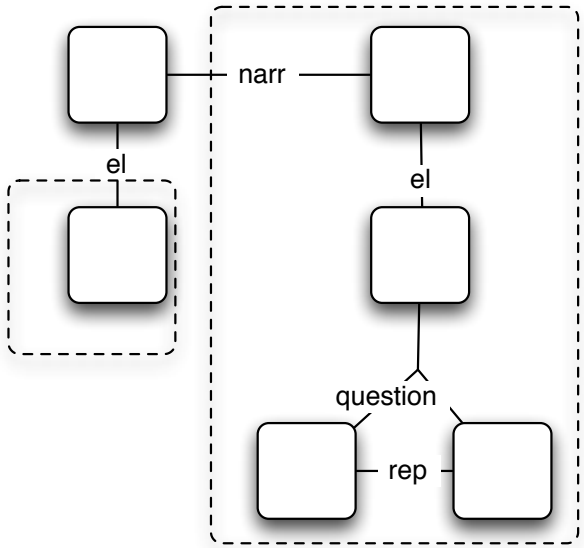
- Two interlocutors, thus two representations of the exchange
  - psychologist : build a representation in spite of pragmatic rules
  - schizophrenic : semantically coherent, but pragmatically incorrect

- Representation :
  - SDRT
  - thematic boxes (dotted box)

Formal representation

- Two interlocutors, thus two representations of the exchange
  - psychologist : build a representation in spite of pragmatic rules
  - schizophrenic : semantically coherent, but pragmatically incorrect

- Representation :
  - SDRT
  - thematic boxes (dotted box)



Context

Two conjectures

Context

Two conjectures

Conjecture 1 : Schizophrenics are logically consistent. Hence the breakings intervene through the construction process of the conversational representation.

Context

Two conjectures

Conjecture 1 : Schizophrenics are logically consistent. Hence the breakings intervene through the construction process of the conversational representation.

Conjecture 2 : Underspecification plays a central role in such breakings.

Context

Two conjectures

Conjecture 1 : Schizophrenics are logically consistent. Hence the breakings intervene through the construction process of the conversational representation.

Conjecture 2 : Underspecification plays a central role in such breakings.

Slogan : “A choice is never definitive !”

Conversation example

B124 Oh yeah (↑) and complicated (↑) and it's really very very complicated (→) **politics**, it's really something when you get into it, **have to win** or else when you lose, well, you're finished (↓)

A125 Yes

B126 **JCD is dead**, L is dead, P is dead uh (...)

A127 So you think **they're dead because they lost** (↑)

B128 No they won but if they're dead, it's **their disease** well it's it's (→)

A129 Yeah it's because they had a disease, it's not because they were in politics (↑)

B130 Yes I mean (→)

A131 Yes you think it's because they were in politics (↑)

B132 Yes, so well yeah there was **C too who committed murder**, uh huh (→) he was there too, the one in B but well (→) it, that, it's because of politics again

Analyse de l'exemple 1

The schizophrenic switch twice from a theme to another one :

Analyse de l'exemple 1

The schizophrenic switch twice from a theme to another one :

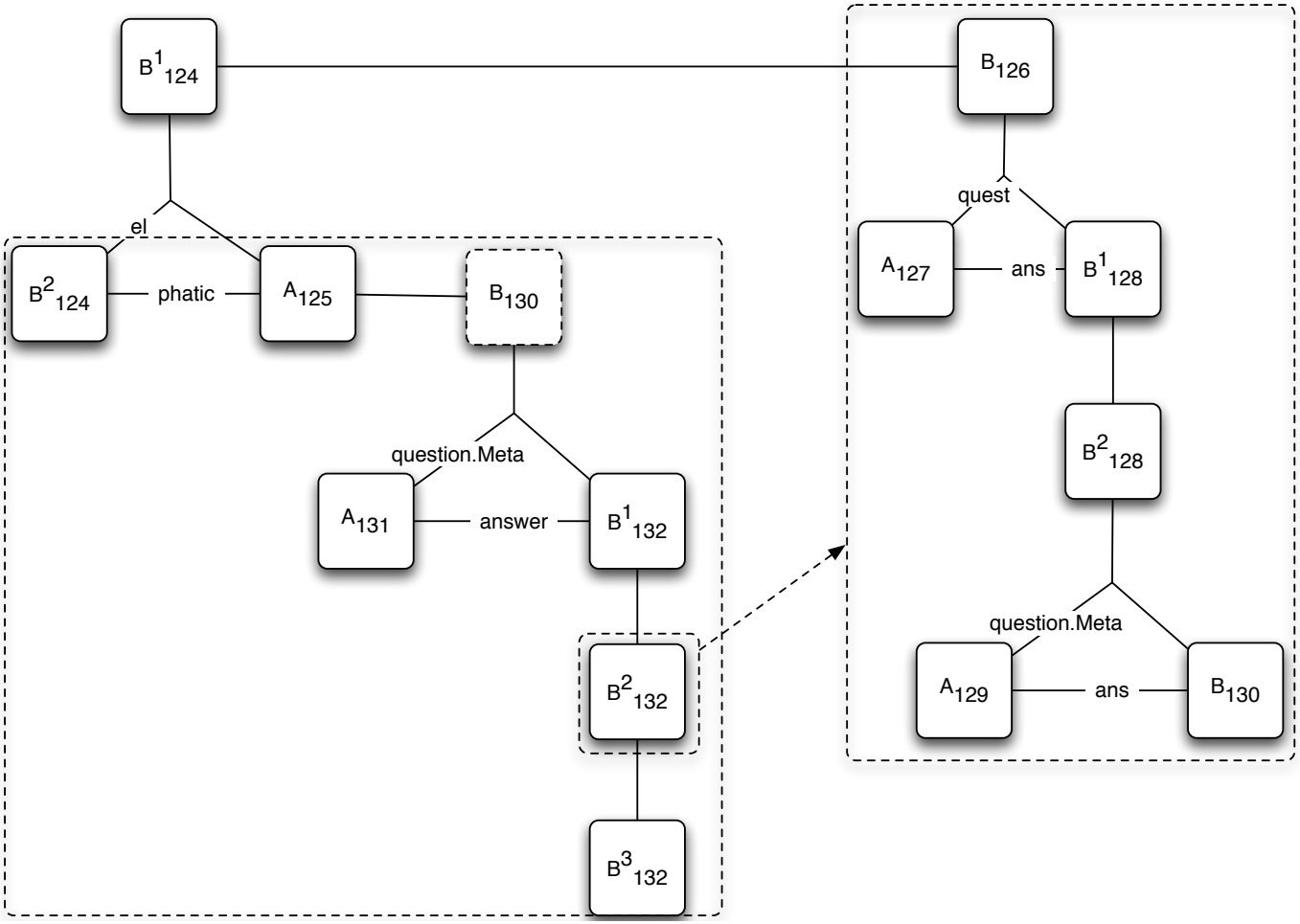
- politic death (symbolic)
- death (literal)

Analyse de l'exemple 1

The schizophrenic switch twice from a theme to another one :

- politic death (symbolic)
- death (literal)

It's clear that the two themes are directly relied, but they express two different realities.
Here, we identify a break of the right frontier constraint.


B1 124
el
B2 124
phatic
A125
B130
question.Meta
A131
answer
B1 132
B2 132
B3 132
B126
quest
A127
ans
B1 128
B2 128
question.Meta
A129
ans
B130

Plan

Semi-Supervised Interview Schizophrenic / Psychologist

- Transcription of the interview
- Neuro-cognitive abilities :
  - `Wechsler Adult Intelligence Scale-III` (measure of IQ)
  - `California Verbal Learning Test` (strategy and cognitive ability)
  - `Trail Making Test` (depreciation of cognitive flexibility and inhibition).
- oculomotor behavior (double eye-trackers)
- Activity of the brain (EEG)

A relatively large corpus

	corpus Town1			corpus Town2			total
	men	women	total	men	women	total	
schizophrenics	15	3	18	20	10	30	48
control	15	8	23	4	4	8	31
total	30	11	41	24	14	38	79

A relatively large corpus

	corpus Town1			corpus Town2			total
	men	women	total	men	women	total	
schizophrenics	15	3	18	20	10	30	48
control	15	8	23	4	4	8	31
total	30	11	41	24	14	38	79

31 575 speechs / 375 000 words

	corpus Town1				corpus Town2			
	nb speechs		nb words		nb speechs		nb words	
S	3 863	11 145	46 859	119 762	4 062	4 433	66 725	79 081
T	7 282		72 903		371		12 356	
$P + S$	3 819	11 517	30 293	138 571	4 098	4 480	33 686	37 842
$P + T$	7 698		108 278		382		4 156	
total		22 662		258 333		8 913		116 923

A corpus difficult to constitute

Traditionally, non-broadcast of corpora

- Inability to access other projects / oldest corpus

A corpus difficult to constitute

Traditionally, non-broadcast of corpora

- Inability to access other projects / oldest corpus

Heavy administrative procedures :

- CPP of the area of the medical institution
  - Finalized description of the protocol
  - description of invasiveness
  - Several month of investigation
  - Take insurance
- CNIL
- The data should not be used for / against the patient.

A corpus difficult to constitute

Patient participation :

- Identifying patients able to participate
- Identify those who agree to participate
- Manage patient concerns
  - patient identification and obtain their agreement (fear of disclosure of biographical data) (lost $\sim$ 45 %)
  - dropout (lost $\sim$ 10 %)

A corpus difficult to constitute

- Transcription
  - Automatically : inconclusive test of different tools for french
  - Hand made : two annotators (including the psychologist)
- Transcription guide

Plan

Automatic Processing ?

- Large corpus
- Limit human interventions

→ automatic identification of disfluencies in the transcripts with `Distagger` [Constant and Dister, 2010]

→ automatic identification of POS and lemmas with `MElt` []

Why ?

- to study their uses (conventionnal *vs* pathological)
- For disfluences, to rebuild more consistent speechs (syntactically)

Distagger

f-score : 95,5 %, précision : 95,3 %, rappel : 95,8 %
[Constant and Dister, 2010]

1. *'euh'*

   (1) *moi ça m'est presque plus euh difficile et euh anti-naturel de parler*

2. *Repeat*

   (2) *j' arrive à être à être concentrée quand il faut faire quelque chose*

3. *Self-corrections*

   (3) *enfin je sais pas trop le les termes*

4. Starters

   (4) *pis progressivement vous av- pouvez travailler sur votre concentration*

Distagger

- Seven tags
  $\{IGN{+}EUH\}$, $\{IGN{+}REP\}$, $\{IGN{+}CORR\}$, $\{IGN{+}FRAG\}$, $\{IGN{+}short\_pause\}$, $\{IGN{+}slot\}$ et $\{IGN{+}speaker\}$

Distagger

- Seven tags
  {*IGN+EUH*}, {*IGN+REP*}, {*IGN+CORR*}, {*IGN+FRAG*},
  {*IGN+short_pause*}, {*IGN+slot*} et {*IGN+speaker*}

$\Rightarrow$ identify speechs and interlocutors

Distagger

- Seven tags
  {*IGN+EUH*}, {*IGN+REP*}, {*IGN+CORR*}, {*IGN+FRAG*}, {*IGN+short_pause*}, {*IGN+slot*} et {*IGN+speaker*}

⇒ identify speechs and interlocutors

⇒ very small quantities (reap. 1 and 5 tags)

Distagger

- Seven tags
  {*IGN+EUH*}, {*IGN+REP*}, {*IGN+CORR*}, {*IGN+FRAG*},
  {*IGN+short_pause*}, {*IGN+slot*} et {*IGN+speaker*}

⇒ identify speechs and interlocutors

⇒ very small quantities (reap. 1 and 5 tags)

- Python Scripts : pre and post-processing

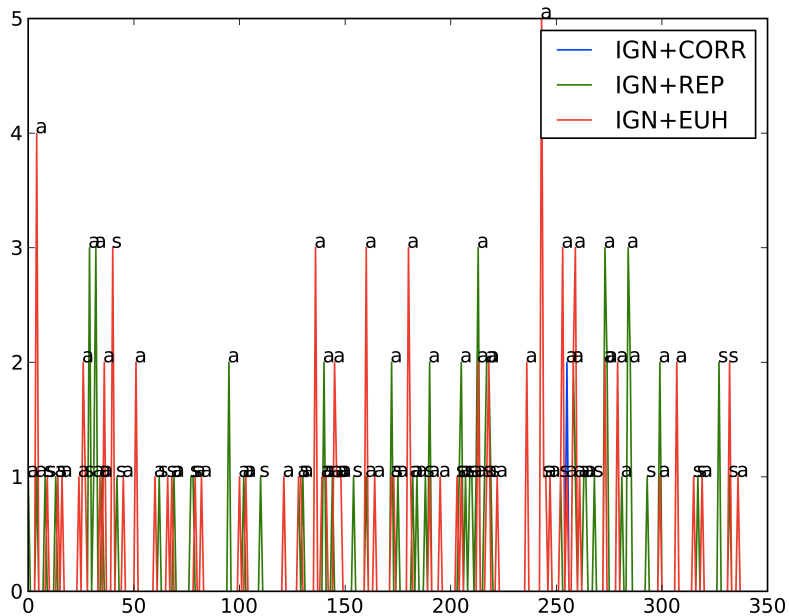
Study of tags

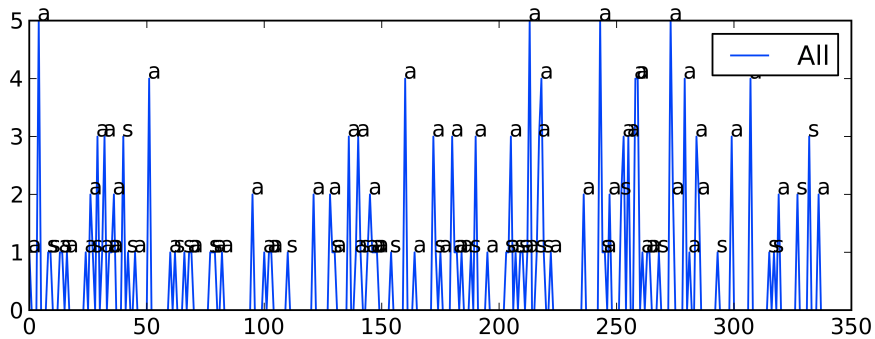
- Some annotations are spread apart :
  'mmh mmh', *'oui oui'*, *'non non'*, *'vous vous'*
- $\sim$ 1 500 code lignes
- Automatic production
  - graphics :
    - by interview : disfluences per speechs (categories and total)
    - by sub-corpus : a surface
  - of values :
    - average of disfluencies per speech and words
    - value of significance (sub-corpus)

Display results

by interview

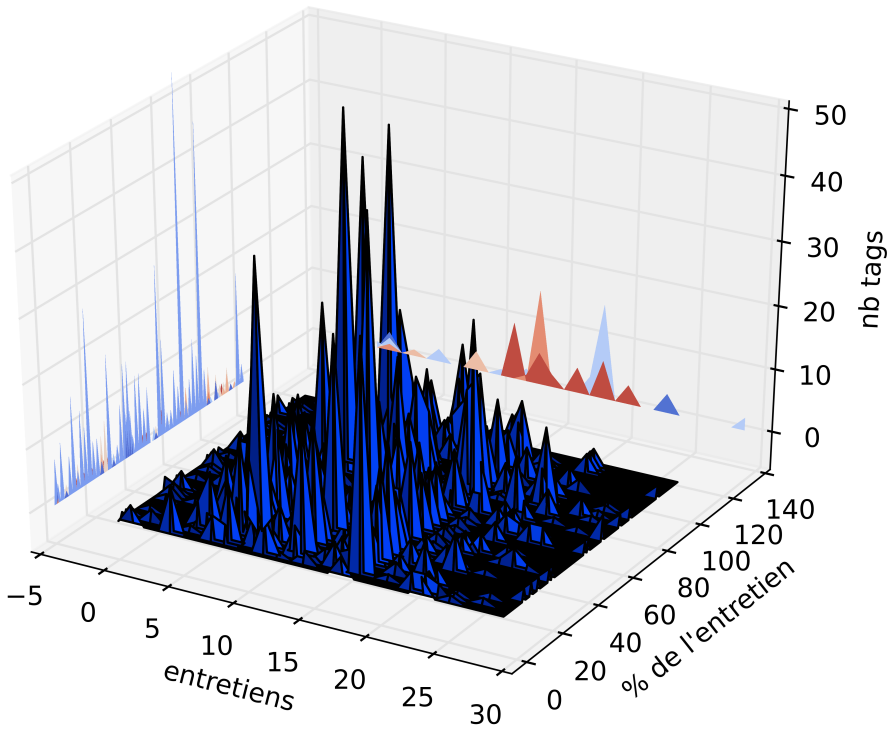
Distribution of disfluencies in an interview

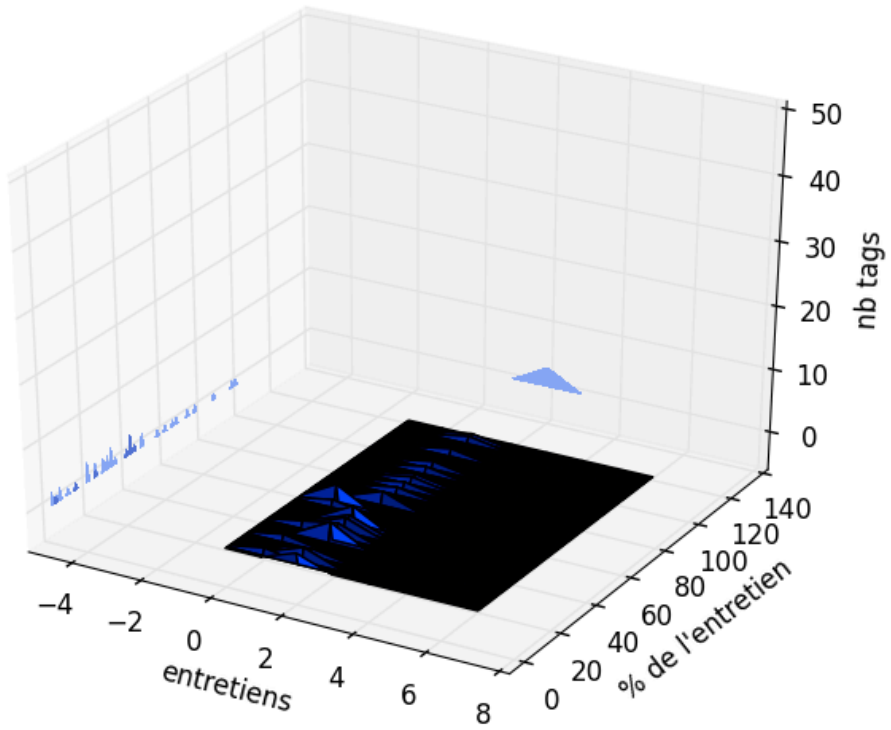




Display results

by sub corpora





Results (% of disfluences)

	S	C	S+C	P+S	P+C	P
			Corpus Town2			
by speech	0, 5417	0, 5589	0, 545	0, 1400	0, 1513	0, 1424
by words	0, 032	0, 0168	0, 0288	0, 0144	0, 0138	0, 0142
			Corpus Town1			
by speech	0, 7117	0, 484	0, 5842	0, 3338	0, 7369	0, 5599
by words	0, 0595	0, 0468	0, 0524	0, 0421	0, 0496	0, 0463

- By speechs : variability
- By words :
  - témoins et psychologue : values of the same order
  - schizophrènes : much higher values

Significance

[de Mareüil et al., 2013]

$$s = \frac{(p_1 - p_2)}{\sqrt{p(1-p)(\frac{1}{n_1} + \frac{1}{n_2})}}$$

where :

$$p = (n_1 p_1 + n_2 p_2)/(n_1 + n_2)$$

Significative values : $> 1,96$

	Town1	Town2
S and Psy	10,6806923083	19,4197596818
C and Psy	0,422898291704	3,23530253756
S and C	10,2827554261	16,0376100956

Bias

- Differences between sub-corpus (different transcription)
  $\Rightarrow$ no overall qualitative assessment
  - `Distagger` the spontaneous speech corpus
    TCOF-POS [Benzitoun et al., 2012] : 4,3 %
- Differences in age and IQ

	age	IQ	studies
Schizophrenics	29 years	95	12,4 years
Control	23 years	103	13 years
Student Test	$p = 0,0058$	$p = 0,0203$	

- Patients under médecine

MElt

- use of Melt tagger [Sagot and Denis, 2012]

  Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art POS tagging with less human effort.

  Form / CAT / Lemma

- Python scripts to pre and post-process the use of Melt (∼400 lignes)

Results

Average on the 41 interviews of Town1

	# cat	# forms	# lemmes	# lemmes/#forms
Schizos	23.667	521.0	409.0	0.785
Control	24.217	628.0	491.0	0.781
Pst (S+C)	23.975	581.0	455.0	0.783

Results

Average on the 41 interviews of Town1

	# cat	# forms	# lemmes	# lemmes/#forms
Schizos	23.667	521.0	409.0	0.785
Control	24.217	628.0	491.0	0.781
Pst (S+C)	23.975	581.0	455.0	0.783

Nothing to say...

Cat	# cat	by words	by speech
Schizos	23.667	0.011	0.132
Control	24.217	0.009	0.083
Pst (S+C)	23.975	0.009	0.104

Cat	# cat	by words	by speech
Schizos	23.667	0.011	0.132
Control	24.217	0.009	0.083
Pst (S+C)	23.975	0.009	0.104

Form	# form	by words	by speech
Schizos	521.0	0.21	2.672
Control	628.0	0.194	2.053
Pst (S+C)	581.0	0.201	2.324

Cat	# cat	by words	by speech
Schizos	23.667	0.011	0.132
Control	24.217	0.009	0.083
Pst (S+C)	23.975	0.009	0.104

Form	# form	by words	by speech
Schizos	521.0	0.21	2.672
Control	628.0	0.194	2.053
Pst (S+C)	581.0	0.201	2.324

Lemmas	# lem	by words	by speech
Schizos	409.0	0.166	2.103
Control	491.0	0.153	1.611
Pst (S+C)	455.0	0.158	1.827

Significance

		Schizo / Temoin	Schizo / Psy	Temoin / Psy
Cat	words	0.698	0.373	-0.324
	speech	1.710	0.860	-0.842
Forms	words	1.496	0.812	-0.689
	speech	-	-	-
Lemmas	words	1.32	0.720	-0.612
	speech	-	-	-

CAT

SCHIZO
(u'V', 6879)
(u'PONCT', 5905)
(u'ADV', 5557)
(u'CLS', 4920)
(u'NC', 4827)
(u'P', 3789)
(u'DET', 3365)
(u'CC', 2230)
(u'PRO', 2195)
(u'ADJ', 1919)
(u'VPP', 1432)
(u'CS', 1349)
(u'VINF', 1289)
(u'CLO', 1266)

PSY
(u'V', 20750)
(u'PONCT', 17687)
(u'ADV', 17629)
(u'NC', 14248)
(u'CLS', 13347)
(u'P', 11230)
(u'DET', 10161)
(u'PRO', 7328)
(u'CC', 7117)
(u'ADJ', 6555)
(u'VINF', 4642)
(u'CS', 4078)
(u'CLO', 3424)
(u'VPP', 2948)

TEMOIN
(u'V', 10723)
(u'PONCT', 9416)
(u'ADV', 8976)
(u'NC', 7743)
(u'CLS', 7441)
(u'P', 5634)
(u'DET', 5436)
(u'CC', 3897)
(u'PRO', 3578)
(u'ADJ', 3377)
(u'CS', 2229)
(u'VINF', 1978)
(u'CLO', 1884)
(u'VPP', 1806)

FORM

SCHIZO	PSY	TEMOIN
(u'...', 2758)	(u'...', 9049)	(u'...', 4310)
(u'.', 1658)	(u'est', 5650)	(u'.', 3090)
(u'euh', 1535)	(u"c'", 4863)	(u'est', 2669)
(u'/', 1413)	(u'euh', 4508)	(u"c'", 2296)
(u'je', 1367)	(u'.', 4377)	(u'euh', 2095)
(u'est', 1292)	(u'ça', 3554)	(u'ça', 1668)
(u"c'", 1159)	(u'de', 2724)	(u'je', 1593)
(u"j'", 993)	(u'/', 2667)	(u'/', 1463)
(u'ça', 983)	(u'et', 2540)	(u'de', 1301)
(u'de', 937)	(u'pas', 2467)	(u'ouais', 1300)
(u'pas', 812)	(u'vous', 1938)	(u'pas', 1233)
(u'ouais', 611)	(u'je', 1754)	(u'et', 980)
(u'et', 606)	(u'en', 1710)	(u'que', 940)
(u'que', 604)	(u'que', 1663)	(u'le', 928)

Lemmas

SCHIZO	PSY	TEMOIN
(u'cln', 3356)	(u'...', 9049)	(u'cln', 4322)
(u'...', 2758)	(u'être', 7882)	(u'...', 4310)
(u'être', 2327)	(u'cln', 6768)	(u'être', 3783)
(u'.', 1658)	(u'ce', 6629)	(u'.', 3090)
(u'ce', 1583)	(u'le', 4943)	(u'ce', 2909)
(u'avoir', 1497)	(u'.', 4377)	(u'le', 2745)
(u'*euh', 1471)	(u'*euh', 4253)	(u'avoir', 2112)
(u'le', 1470)	(u'avoir', 3601)	(u'*euh', 1936)
(u'*/', 1211)	(u'cela', 3554)	(u'cela', 1670)
(u'de', 1099)	(u'de', 3281)	(u'de', 1531)
(u'cela', 987)	(u'un', 2941)	(u'un', 1466)
(u'un', 961)	(u'et', 2913)	(u'*/', 1327)
(u'pas', 819)	(u'pas', 2474)	(u'que', 1286)
(u'que', 770)	(u'que', 2461)	(u'*ouais', 1277)

Plan

Traditional anonymization

- Identification and substitution of named entities
  - efficient but not operational tool [Grouin and Zweigenbaum, 2013]

  → Python scripts

- Significant human intervention
  - 10 categories
  - Identification with uppercase
  - manual check

spk1 So you live on Town1 ↑
spk2 Town2
spk1 Town2 Okay so I am not from here, it's not very far

- Ability to add " beep " on the soundtrack

Normalization of the corpus

- Normalization : around thirty regular expressions
- Identification of speechs with unique number
- Status of the subject (patient *vs* Control) in the structure
- Psychologist taken into account for the dynamics of the interaction

spk1 Et donc euh j' avais j' ai pendant trois trois quatre ans **j' avais commencé des études** j' ai fait un peu différentes choses parce que

...

spk1 Euh **dans une école d' ingénieur à Town1 dans dans le nord** euhh et donc euhh euhh ouais donc j' ai je c' est là où j' ai commencé à être malade en fait juste [...]

spk1 donc du coup ben là c' est je j' ai j' ai repéré deux trois le le c' était quand même assez stressant euh **la la prépa**

spk2 Mmh mmh

spk1 donc euh donc du coup ouais euh et bon pour euh en ce qui concerne les études donc du coup après j' ai j' ai arrêté le le le le l' école d' ingénieur enfin la prépa **je suis revenue à Town2**

spk2 Mmh mmh

spk1 **j' ai fait euh une une une fac de de maths** je suis allé en fac de maths

spk1 à à avoir des délires de persécution tout ça j' ai commencé à à penser à la schizophrénie Euhh mais bon en même temps **juste avant de au lycée je je faisais quand même une grosse dépression**

spk1 Donc euh et donc euh du coup euhh ouais donc euh **à Town2** pareil y avait encore la la dépression qui s' installait euh j' étais dans **un appart en fait j' étais place Lieux3**

...

spk1 et c'était très très gênant

spk2 Ben **c'est le centre de Town2**

Impossibility of anonymization

- Task with small context : *randomise* speechs
- Inability to anonymize the history and geography
  - limit the number stakeholders on the resource

Conclusion

- Production of a standardized resource with metadata
- Pathological use of disfluencies with tools from NLP.
- Formal approaches of pathological conversations with SDRT

- Respect the anonymity of people without sacrificing scientific issues
- Identification of conclusions consequences
- Correlate these indices with other measures (psycho-cognitive tests, eye-trackers and EEG)


Amblard, M. and Fort, K. (2014).
Étude quantitative des disfluences dans le discours de schizophrènes : automatiser pour limiter les biais.
In *TALN - Traitement Automatique des Langues Naturelles*, pages 292–303, Marseille, France.
5, 6


Benzitoun, C., Fort, K., and Sagot, B. (2012).
TCOF-POS : un corpus libre de français parlé annoté en morphosyntaxe.
In *Traitement Automatique des Langues Naturelles (TALN)*, pages 99–112, Grenoble, France.
42


Chaika, E. (1974).
A linguist looks at “schizophrenic” language.
Brain and Language, 1(3) :257–276.
2


Constant, M. and Dister, A. (2010).

Automatic detection of disfluencies in speech transcriptions.
In Pettorino, M., Giannini, A., Chiari, I., and Dovetto, F., editors, *Spoken Communication*, volume 1, pages 259–272. Cambridge Scholars Publishing.
31, 32


de Mareüil, P. B., Adda, G., Adda-Decker, M., Barras, C., Habert, B., and Paroubek, P. (2013).
Une étude quantitative des marqueurs discursifs, disfluences et chevauchements de parole dans des interviews politiques.
TIPA. Travaux interdisciplinaires sur la parole et le langage [En ligne], 29.
mis en ligne le 19 décembre 2013, consulté le 14 février 2014.
41


Fromkin, V. A. (1975).
A linguist looks at “a linguist looks at ‘schizophrenic language’”.
Brain and Language, 2(0) :498 – 503.
2


Grouin, C. and Zweigenbaum, P. (2013).
Automatic de-identification of french clinical records : Comparison of rule-based and machine-learning approches.
In *Stud Health Technol Inform, Proc of MEDINFO*, volume 192, pages 476–80, Copenhagen, Denmark.
54


Musiol, M. and Trognon, A. (1996).
L'accomplissement interactionnel du trouble schizophrénique.
Raisons Pratiques 7, pages 179–209.
2

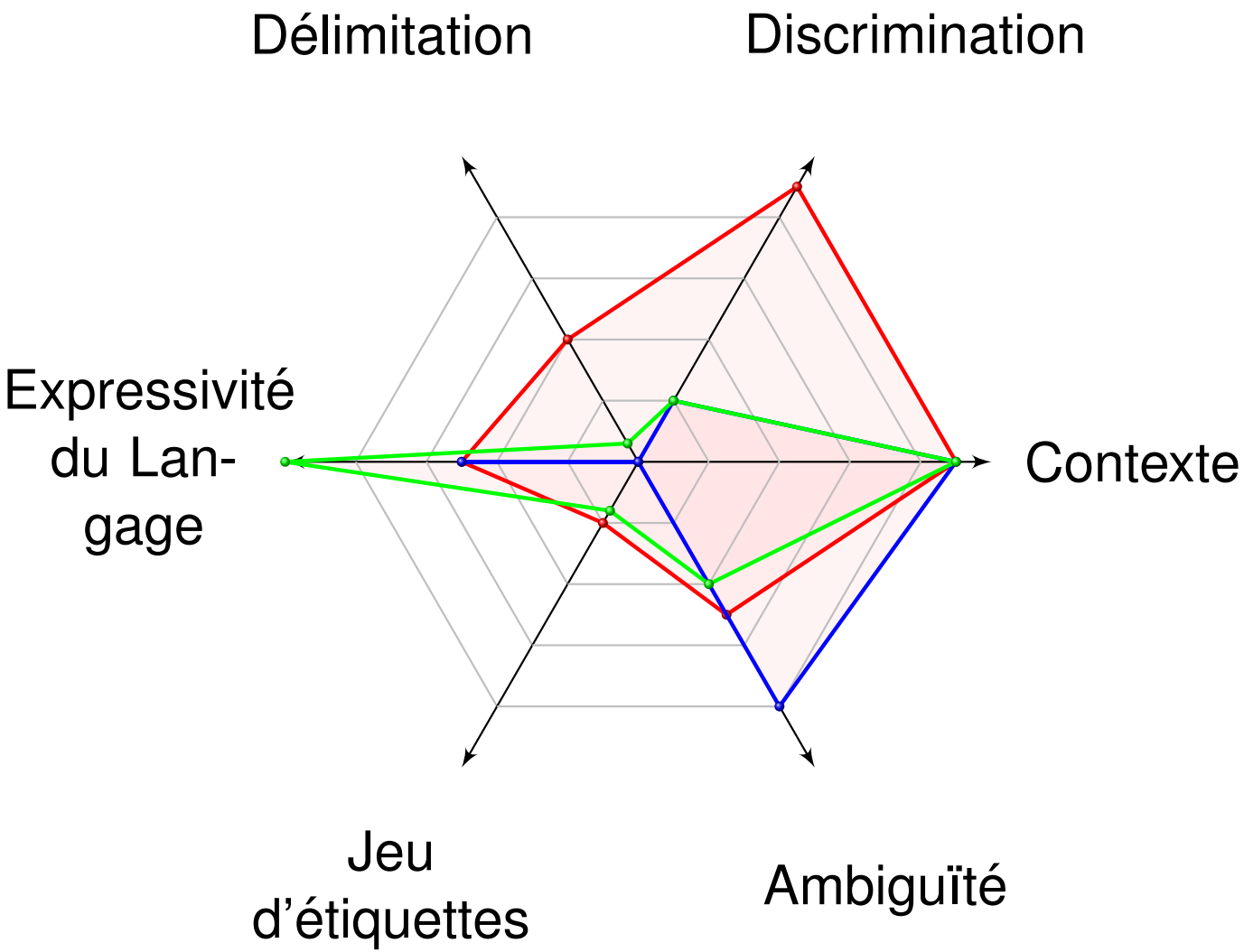

Rebuschi, M., Amblard, M., and Musiol, M. (2014).
Using SDRT to analyze pathological conversations. Logicality, rationality and pragmatic deviances.
In *Interdisciplinary Works in Logic, Epistemology, Psychology and Linguistics : Dialogue, Rationality, and Formalism*, Logic, Argumentation & Reasoning, pages 343–368. Springer.
5, 6


Verhaegen, F. (2007).
Psychopathologie cognitive des processus intentionnels schizophréniques dans l'interaction verbale.
PhD thesis, Université Nancy 2, France.

Modéliser une campagne d'annotation

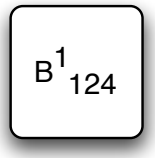
- Éviter les biais
- Définir les mesures des accords inter-annotateurs
- Identifier les outils techniques (et donc la documentation et l'accès à la documentation)
- Prévoir les coûts
- Définir la structure des campagnes

Complexité de l'annotation

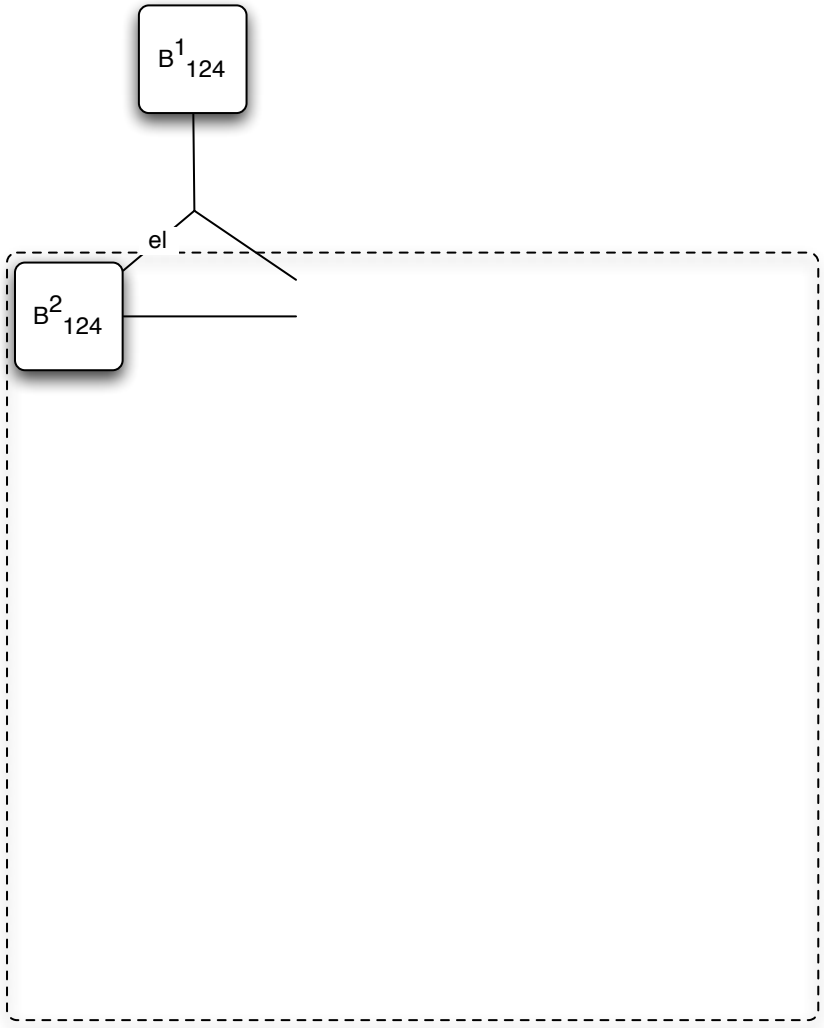


Discontinuités Syntaxe S-DRT

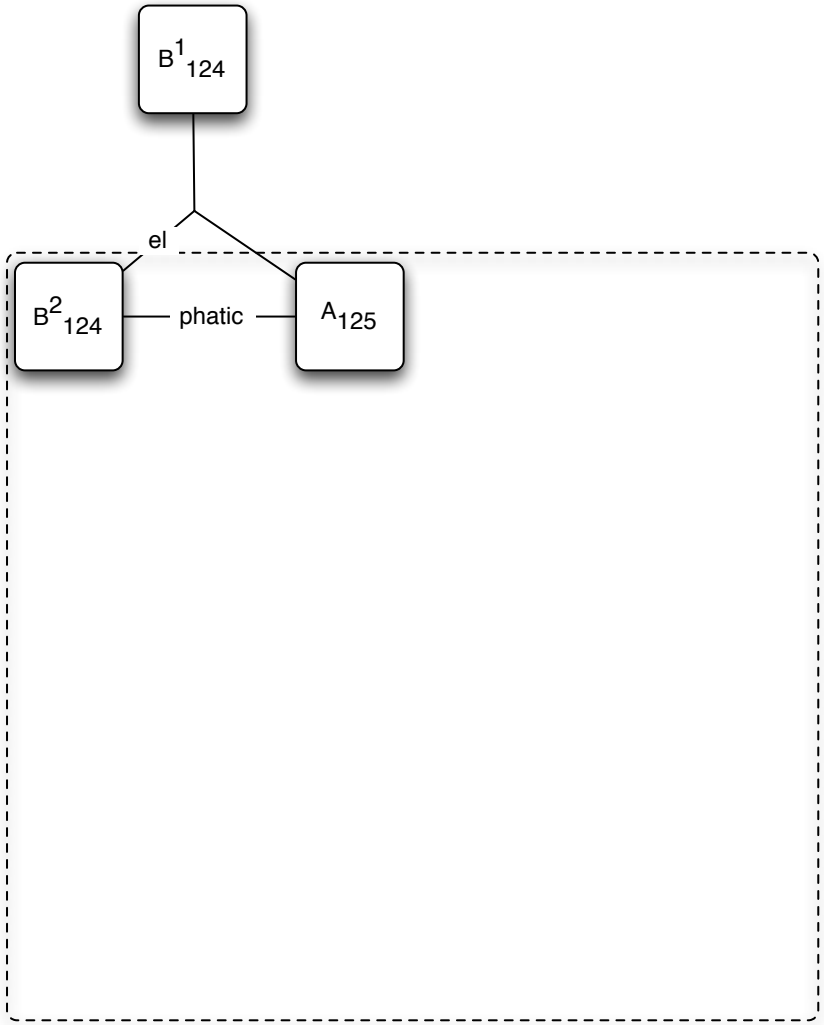
(B124) Oh yeah (↑) and complicated (↑) and it's really very very complicated (→)



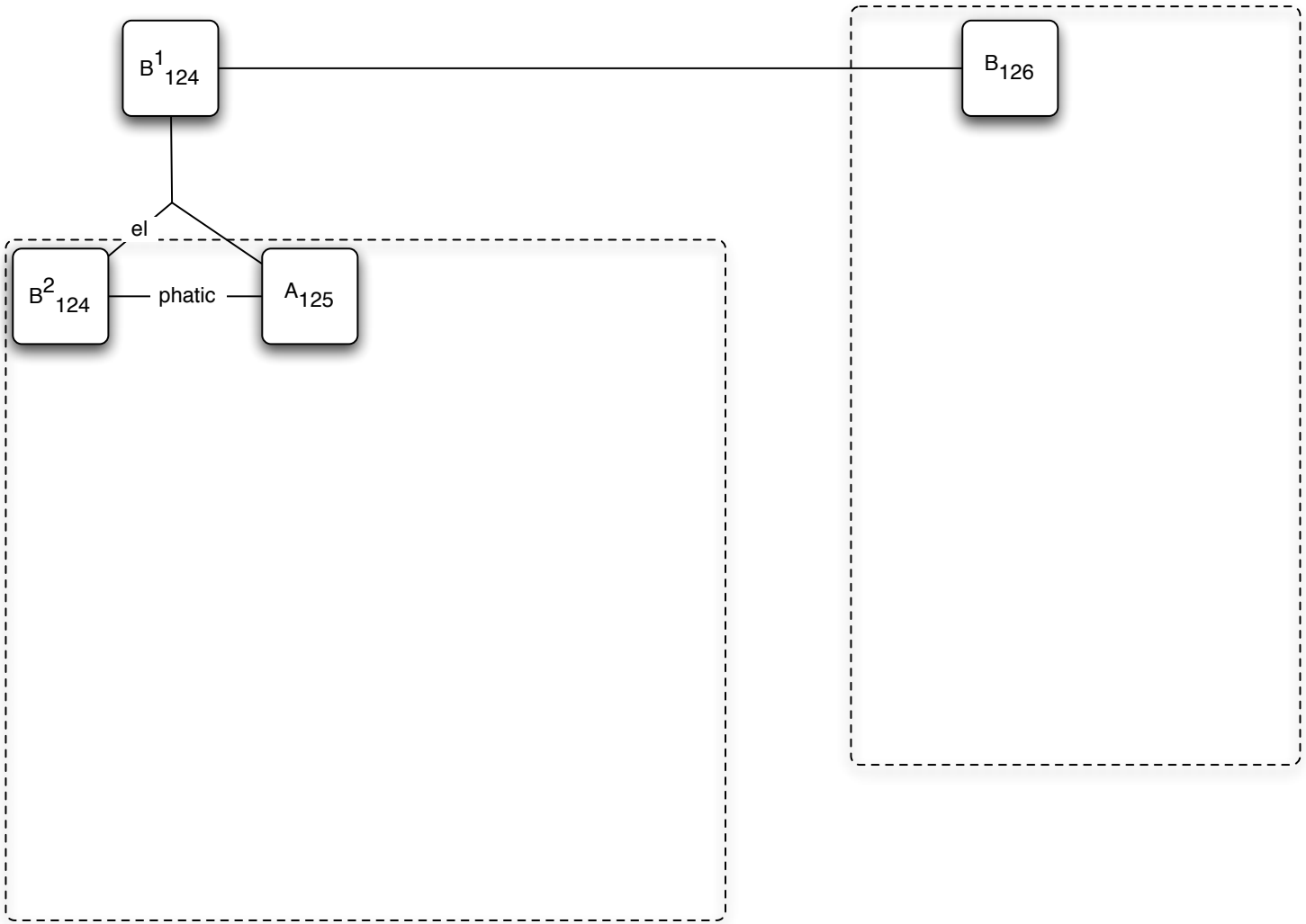
politics, it's really something when you get into it, have to win or else when you lose, well, you're finished ($\downarrow$)



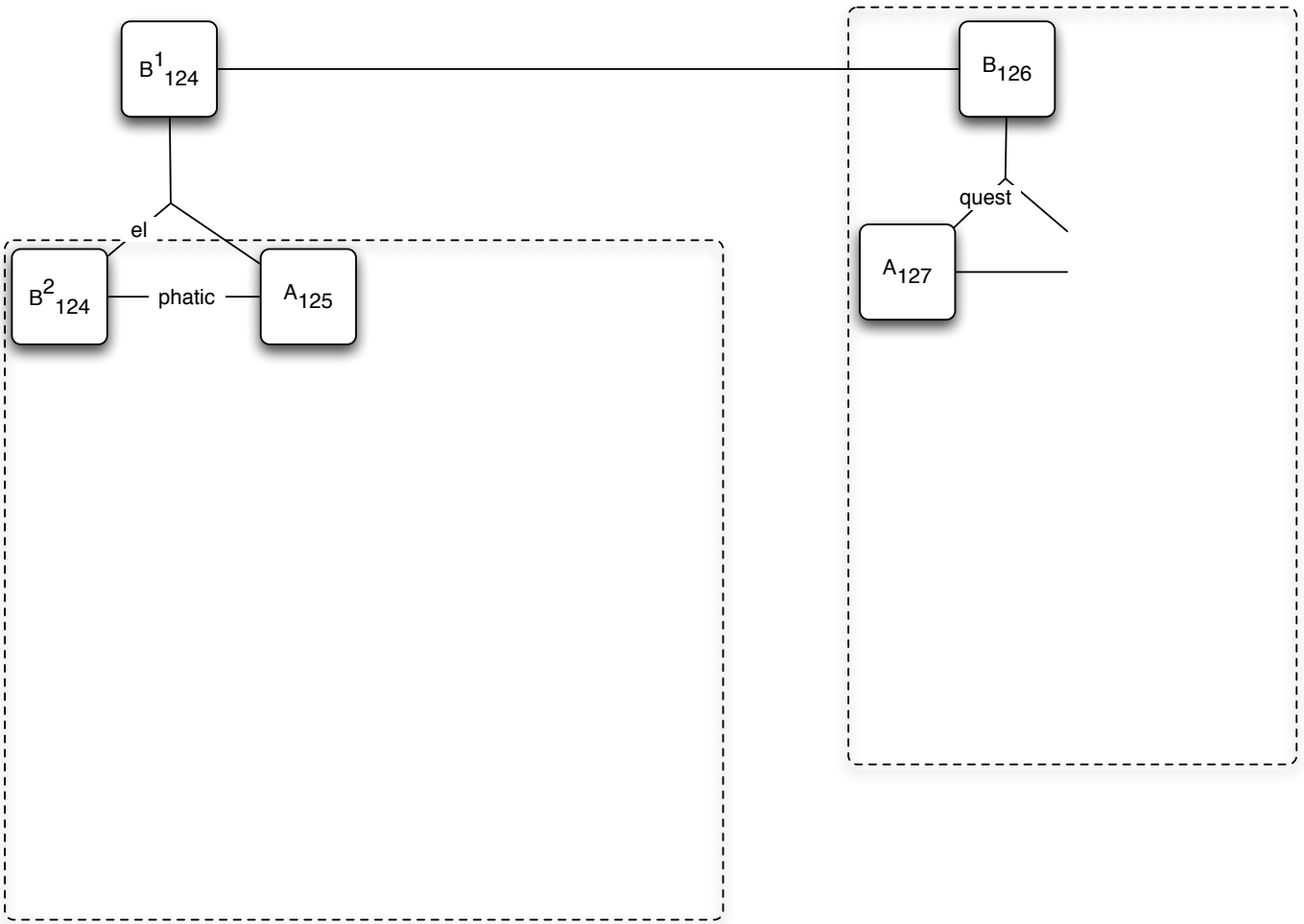
(A125) Yes



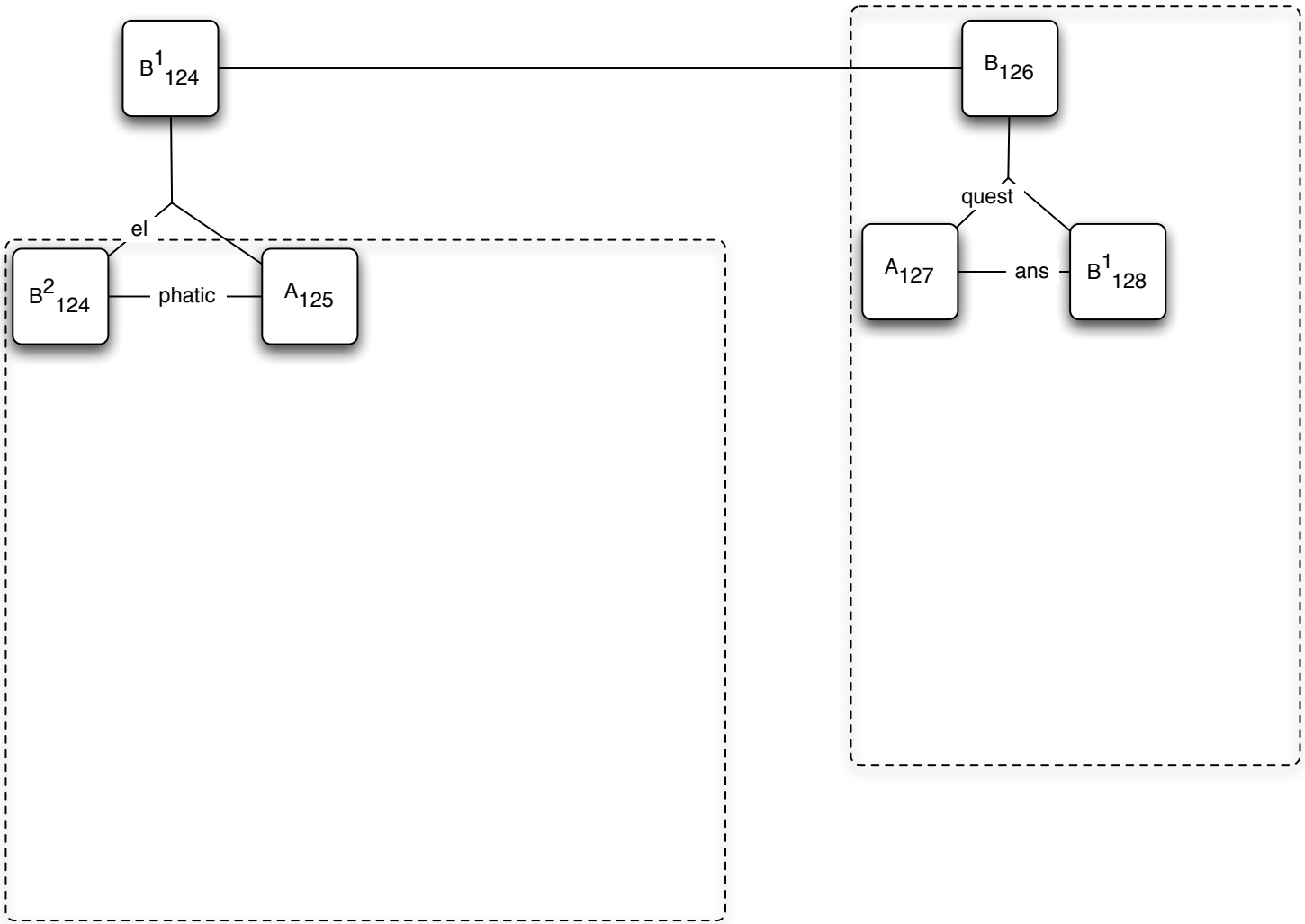
(B126) JCD is dead, L is dead, P is dead uh (...)



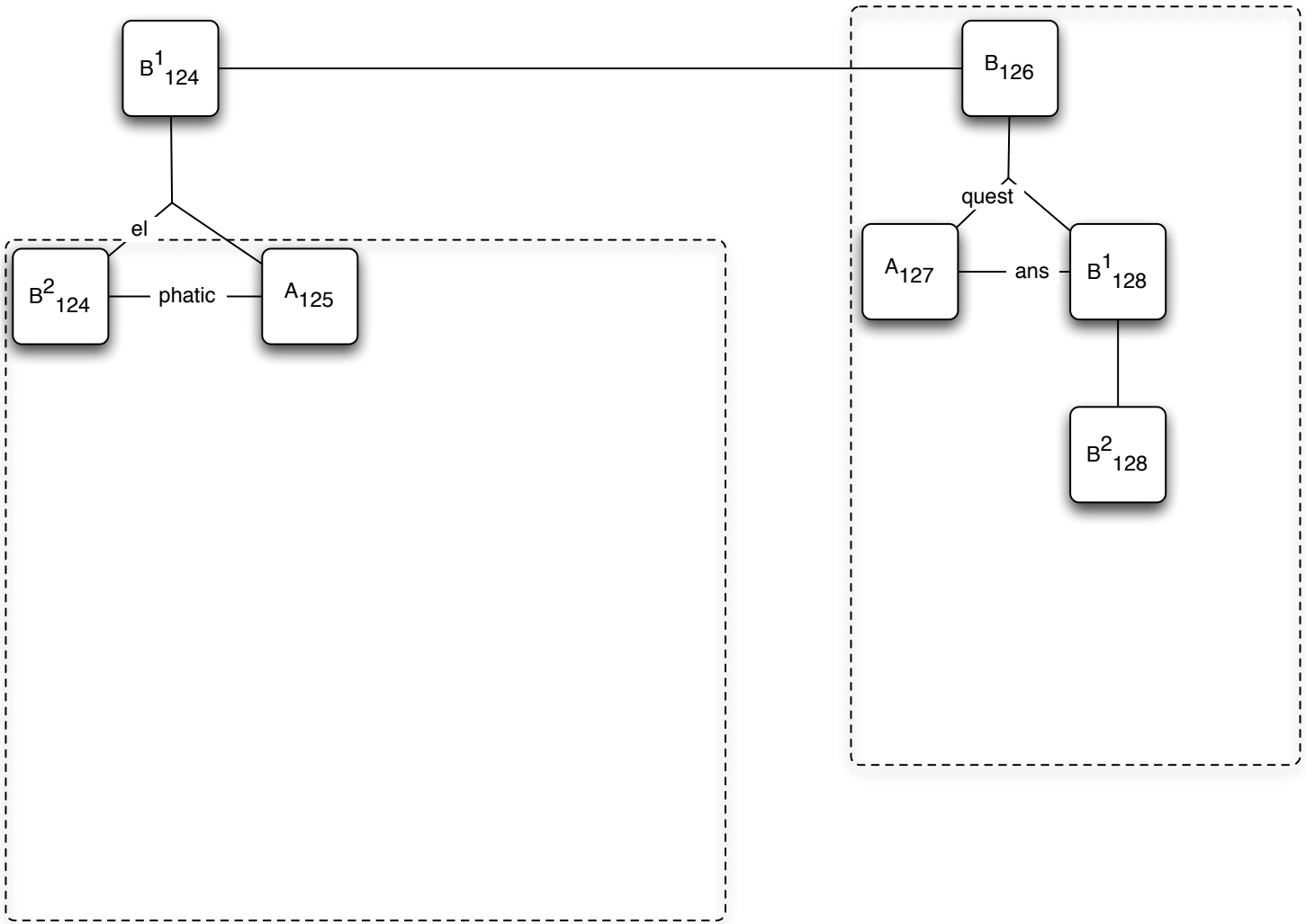
(A127) So you think they're dead because they lost (↑)



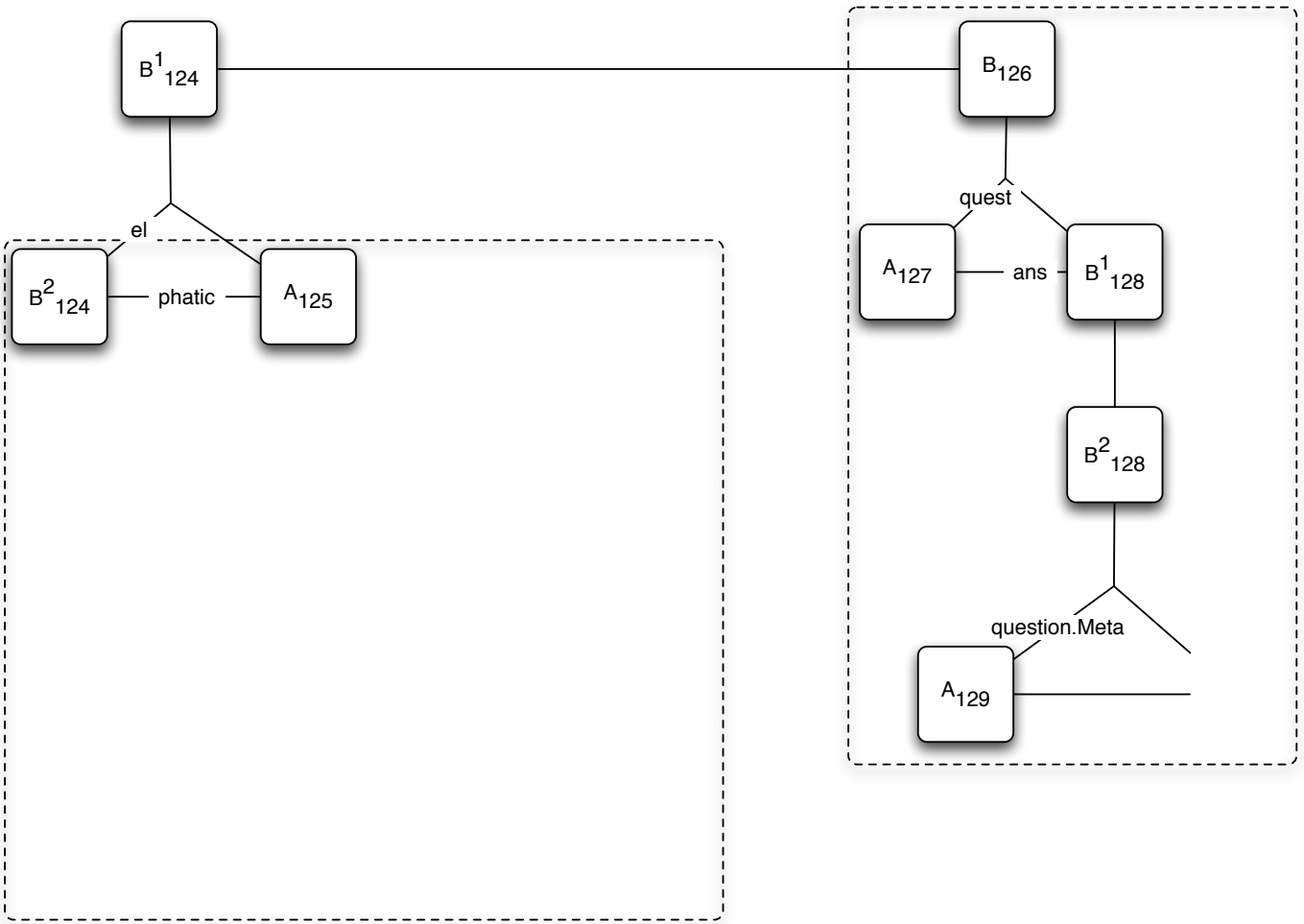
(B128) No they won but if they're dead, it's their disease well it's it's (→)



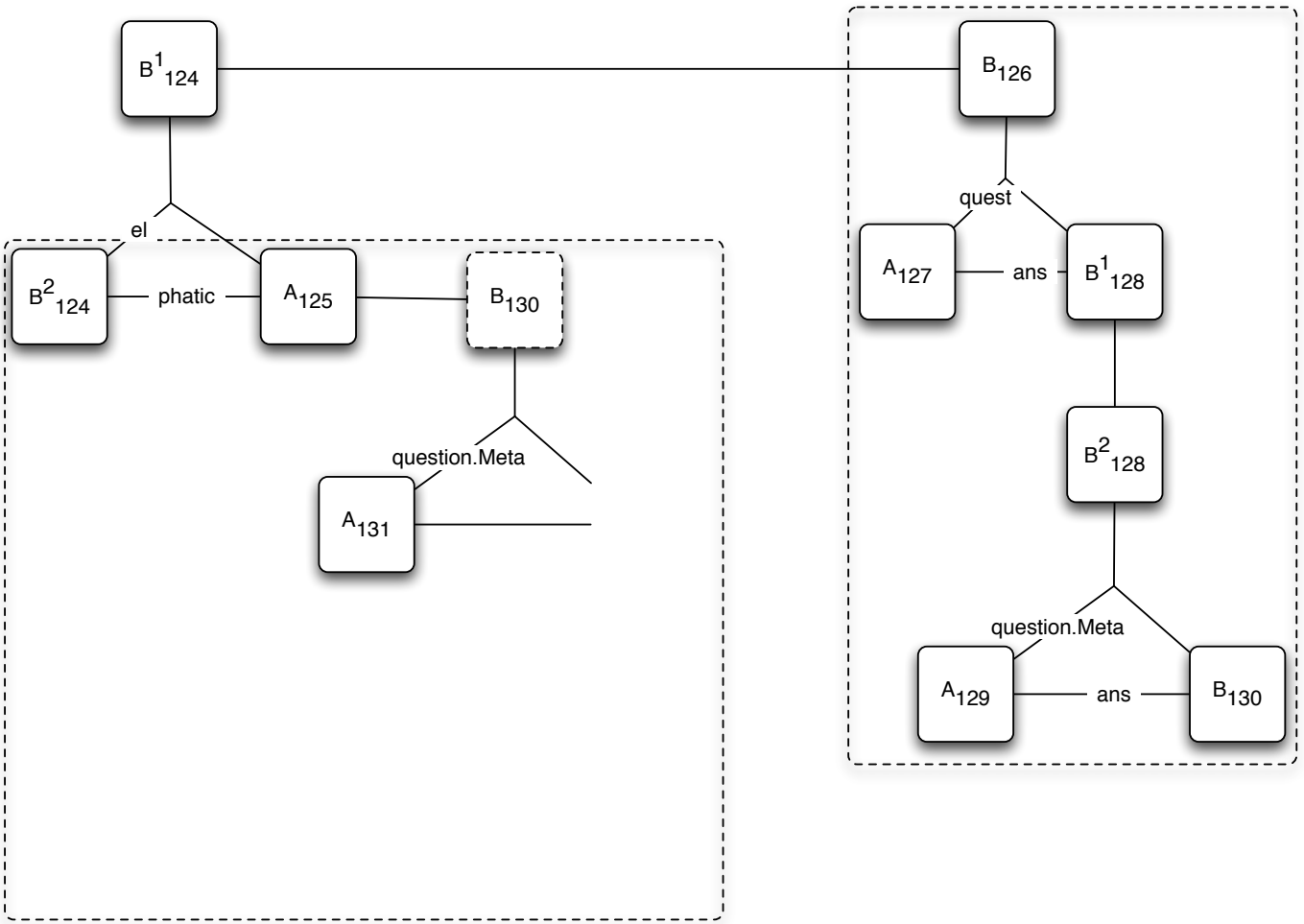
(B128) No they won but if they're dead, it's their disease well it's it's (→)



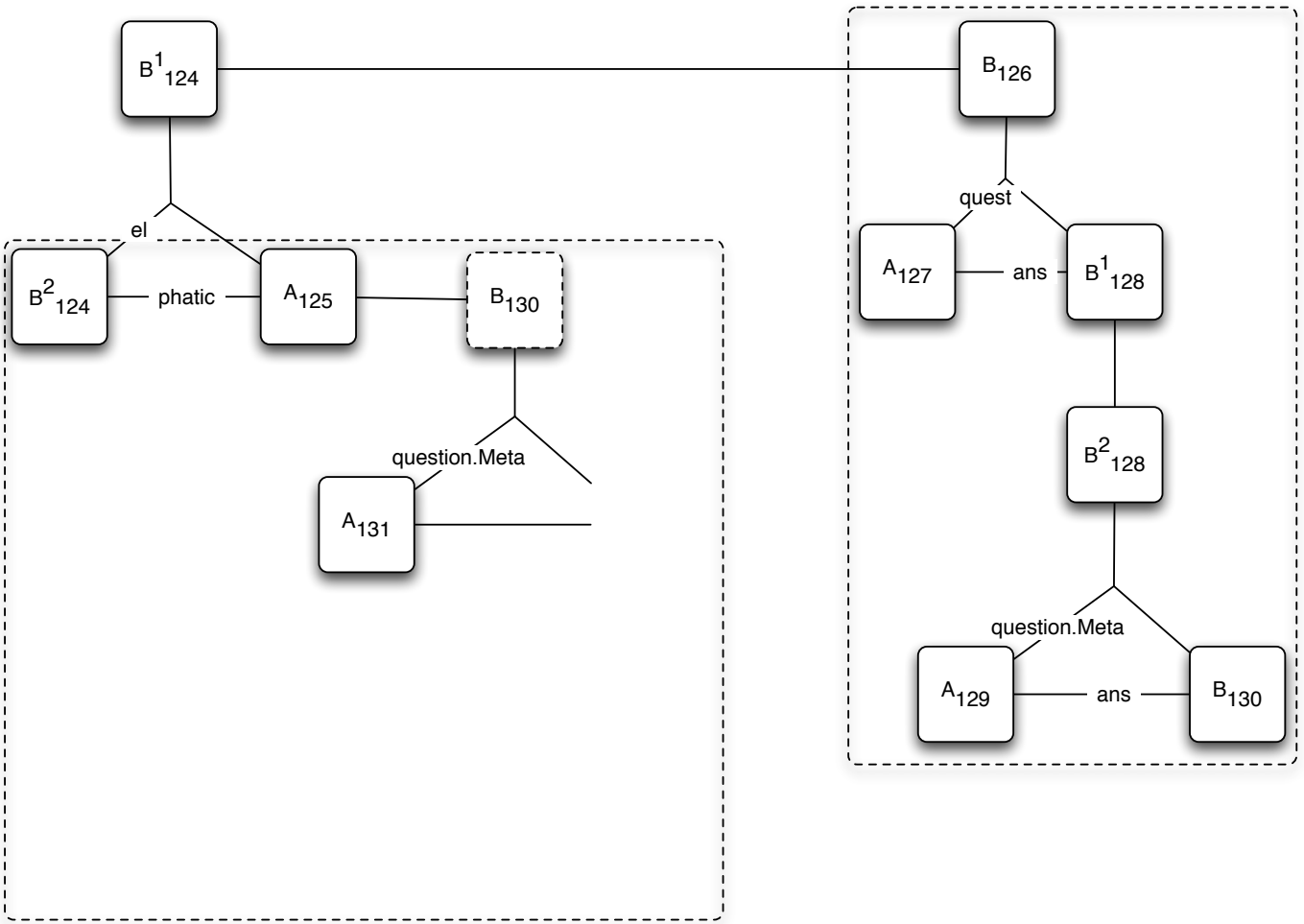
(A129) Yeah it's because they had a disease, it's not because they were in politics (↑)



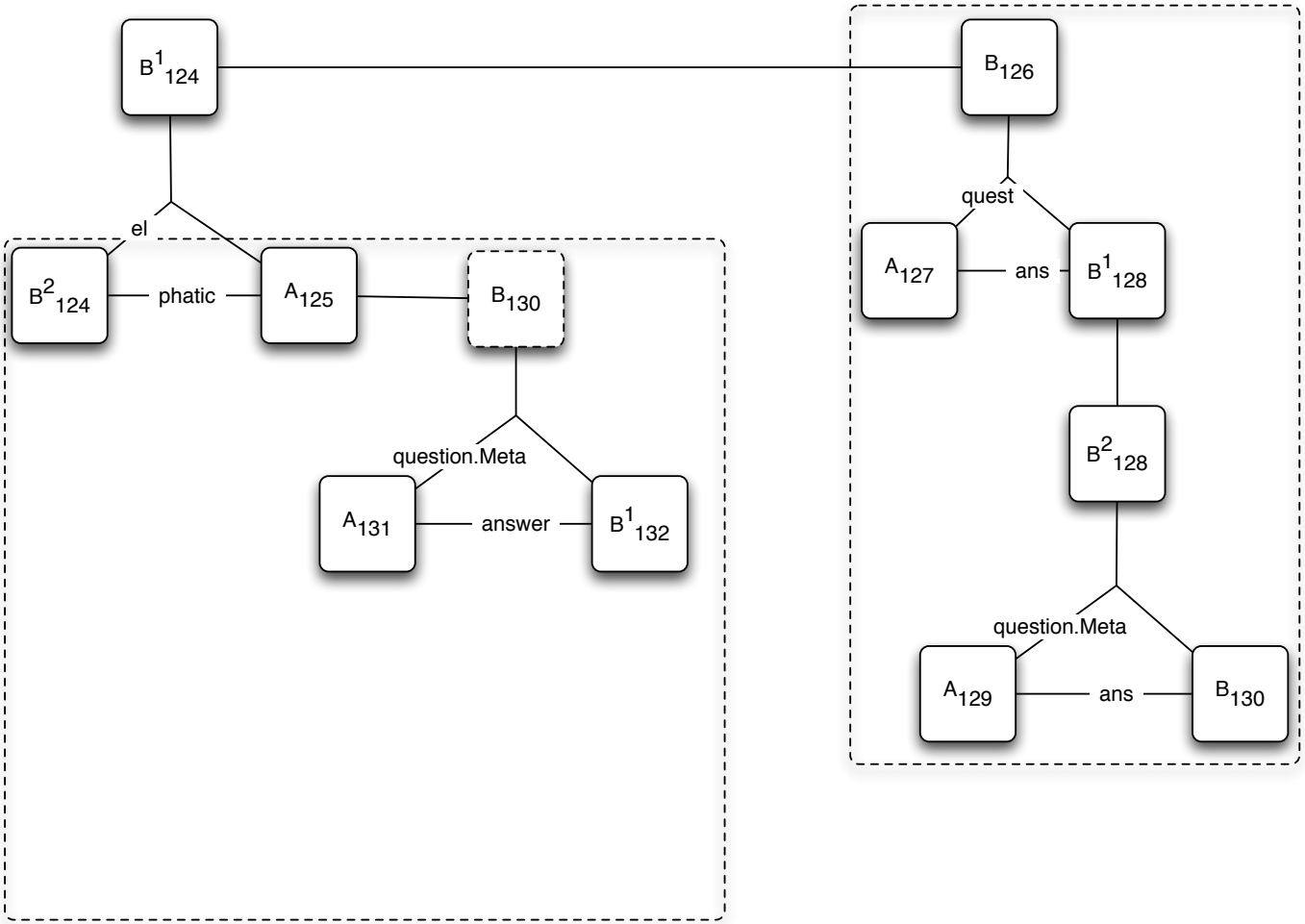
(B130) Yes I mean (→)



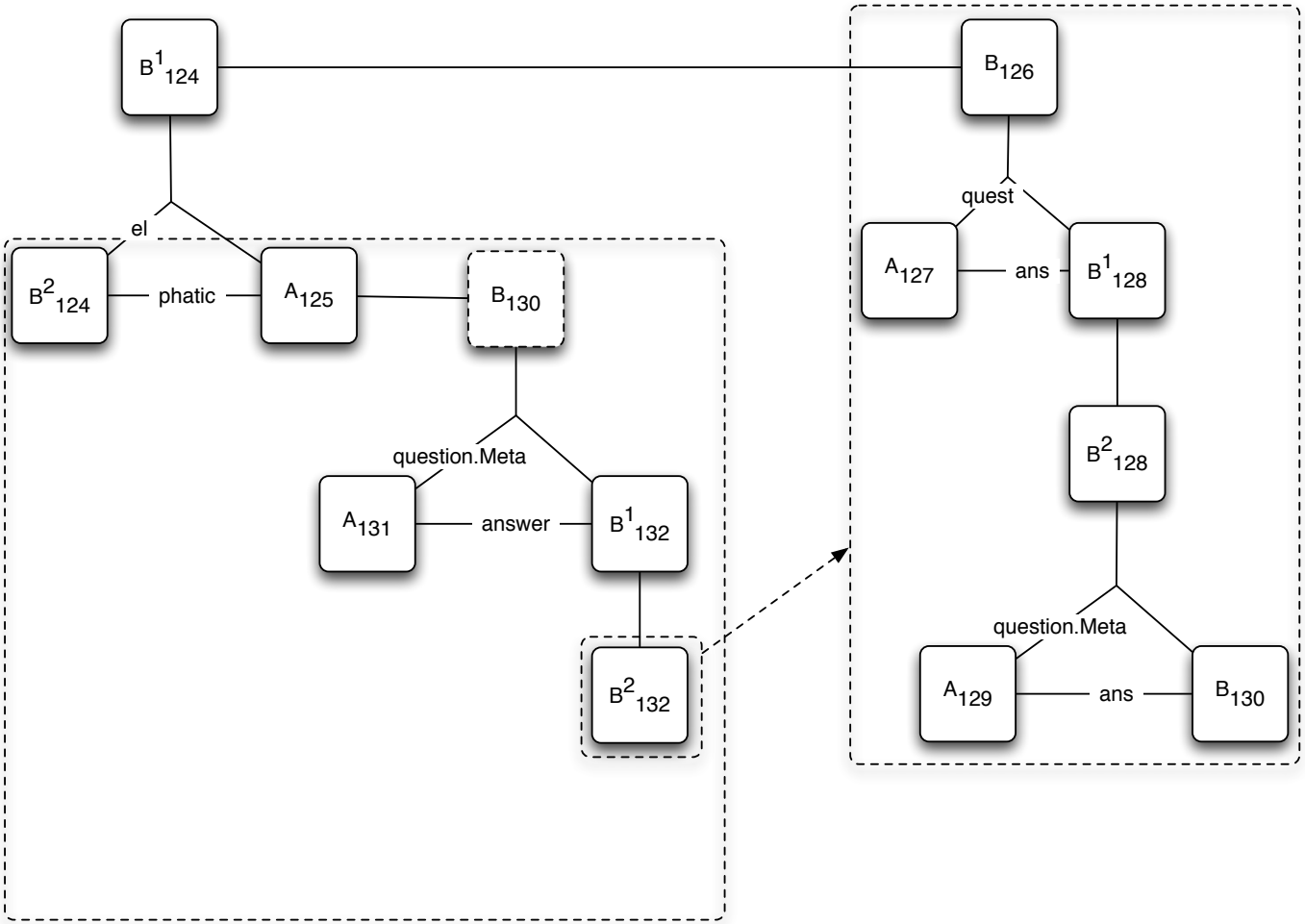
(A131) Yes you think it's because they were in politics (↑)



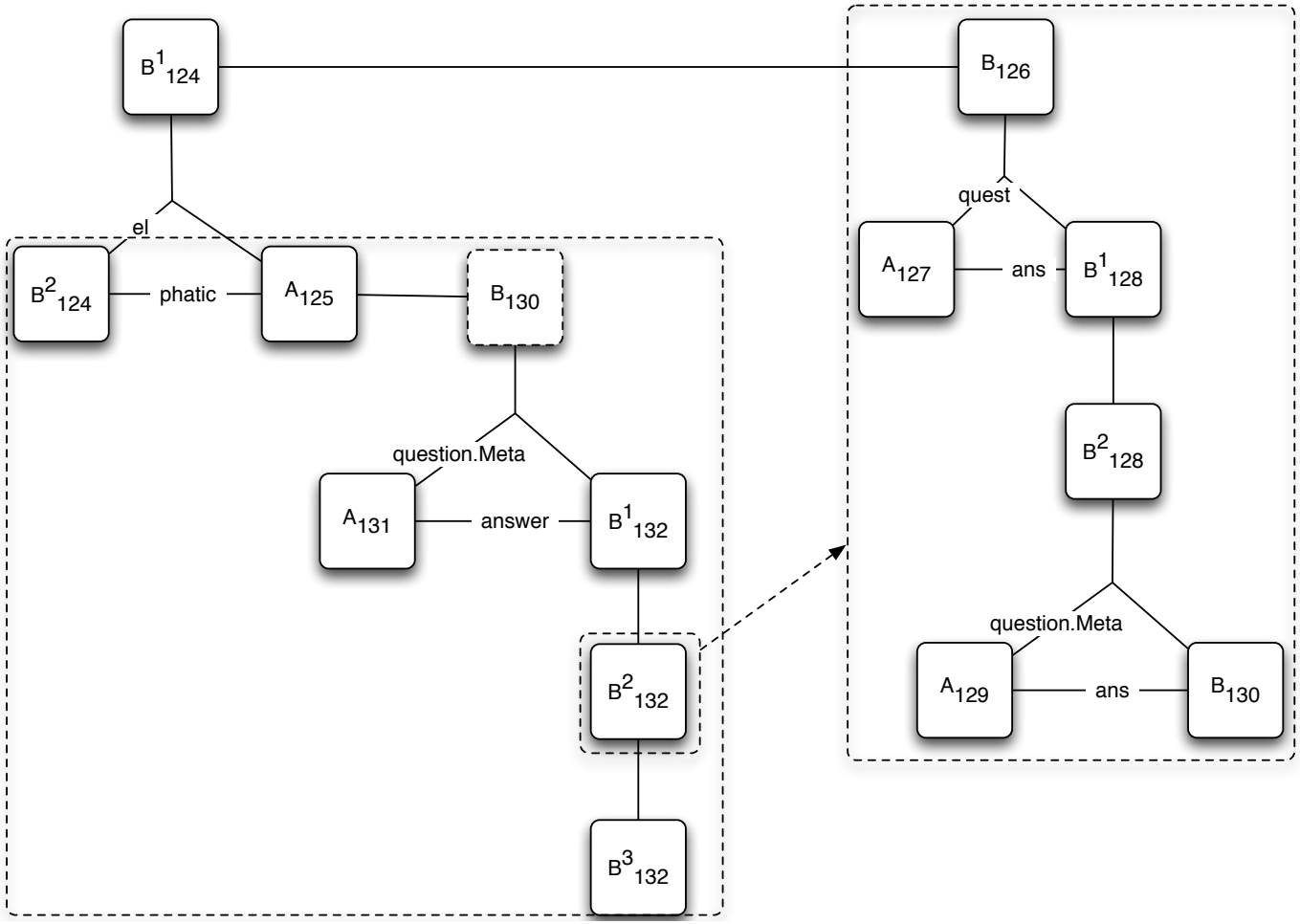
(B132) Yes, so well yeah there was C too who committed murder, uh huh (→) he was there too, the one in B but well (→) it, that, it's because of politics again



(B132) Yes, so well yeah there was C too who committed murder, uh huh (→) he was there too, the one in B but well (→) it, that, it's because of politics again



(B132) Yes, so well yeah there was C too who committed murder, uh huh (→) he was there too, the one in B but well (→) it, that, it's because of politics again



Remontée à travers la structure