L'impossibilité de l'anonymat dans le cadre de l'analyse du discours

Maxime Amblard[1,4,7,8,9] Karën Fort[5,6] Michel Musiol[2,4,7,9]
Manuel Rebuschi[3,4,7,9]

22 Novembre 2014

Journée ATALA : Éthique et TAL

1. LORIA - UMR 7503
2. ATILF - UMR 7118
3. LHS-AHP - UMR 7117
4. MSH-Lorraine - USR 3261
5. STIH - EA 4509
6. Université Paris Sorbonne
7. Université de Lorraine
8. CNRS
9. Inria Nancy Grand Est

Plan

Plan

Motivation

- Étude linguistique de la pathologie mentale [Chaika, 1974] et [Fromkin, 1975]
- *Discontinuités pragmatiques* dans l'accomplissement de l'interaction verbale, [Musiol and Trognon, 1996]
- Usage pathologique de la planification du discours chez les schizophrènes paranoïdes, [Verhaegen, 2007]

SLAM - Schizophrénie et Language : Analyse et modélisation

- Constitution d'une ressource linguistique sur la pathologie mentale
- Études épistémologique et philosophique (norme, folie, rationalité)
- Identifier ces usages par l'utilisation de :
  - modèles formels (type SDRT)
  - outils et méthodes du TAL

SLAM - Schizophrénie et Language : Analyse et modélisation

- Constitution d'une ressource linguistique sur la pathologie mentale
- Études épistémologique et philosophique (norme, folie, rationalité)
- Identifier ces usages par l'utilisation de :
  - modèles formels (type SDRT)
  - outils et méthodes du TAL
- [Rebuschi et al., 2014] : des corrélations explicites
- [Amblard and Fort, 2014] : usage pathologique des disfluences

Rejouer les ambiguïtés linguistiques

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

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**

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

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

Entretien semi-dirigé schizophrène/psychologue

- Contenu de l'entretien (transcription)
- Capacités neuro-cognitives :
  - `Wechsler Adult Intelligence Scale-III`
    (mesure du quotient intellectuel, ou QI)
  - `California Verbal Learning Test`
    (capacité cognitive et de stratégie)
  - `Trail Making Test`
    (dépréciation de la flexibilité cognitive et de l'inhibition).
- Comportement oculomoteur (double système d'*eye-tracker*)
- Activité de l'encéphale (EEG)

Un corpus relativement important

	corpus Ville1			corpus Ville2			total
	hommes	femmes	total	hommes	femmes	total	
schizophrènes	15	3	18	20	10	30	48
témoins	15	8	23	4	4	8	31
total	30	11	41	24	14	38	79

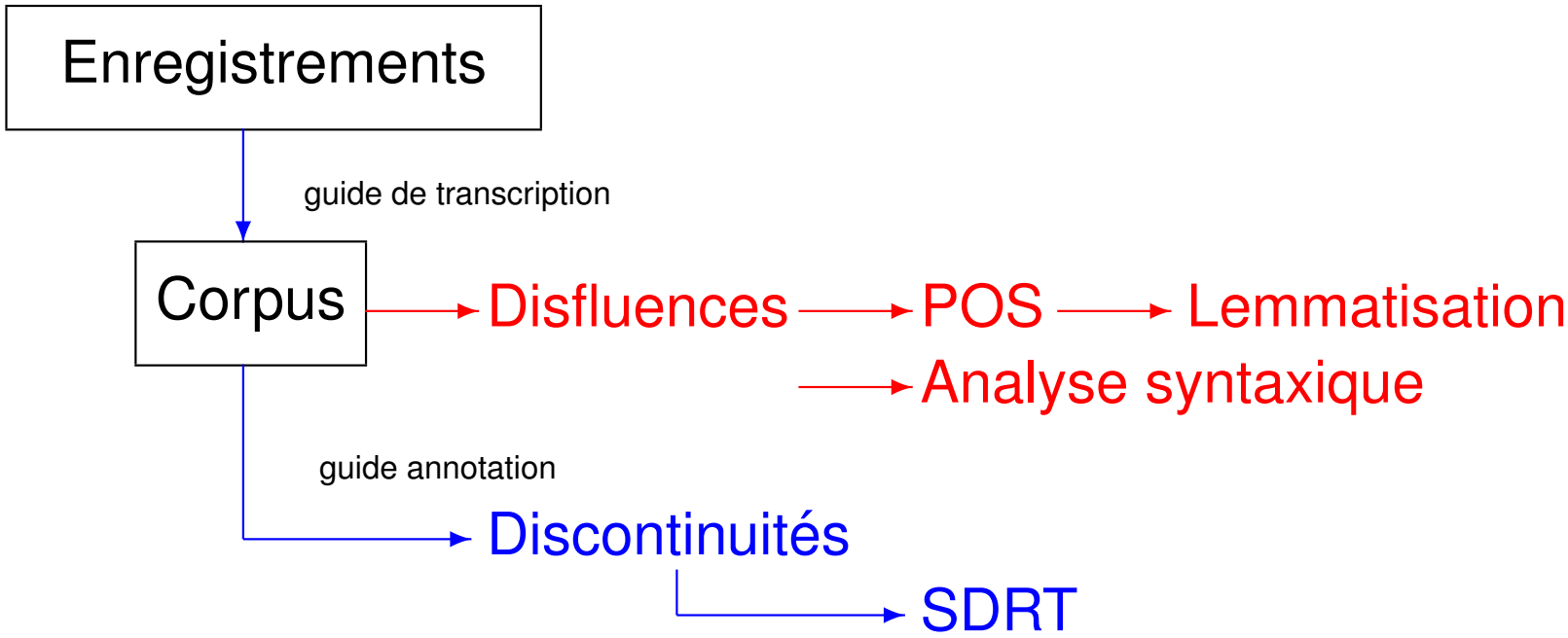
Un corpus relativement important

	corpus Ville1			corpus Ville2			total
	hommes	femmes	total	hommes	femmes	total	
schizophrènes	15	3	18	20	10	30	48
témoins	15	8	23	4	4	8	31
total	30	11	41	24	14	38	79

31 575 tours de parole / 375 000 mots

	corpus Ville1				corpus Ville2			
	nb tours de parole		nb de mots		nb tours de parole		nb de mots	
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

Contexte du projet SLAM



Tradition propriétaire des corpus

Difficulté de mise en perspective de ce type de travaux :

- Impossibilité d'accéder à des corpus d'autres projets / plus anciens

Tradition propriétaire des corpus

Difficulté de mise en perspective de ce type de travaux :

- Impossibilité d'accéder à des corpus d'autres projets / plus anciens

$\Rightarrow$ Nécessité de transférer les bonnes pratiques de la linguistique en matière de gestion des corpus vers les autres disciplines.

Plan

Un corpus difficile à constituer

Démarches administratives lourdes :

- CPP de la région de l'institution médicale
  - Description finalisée du protocole
  - Intrusion du protocole
  - Plusieurs mois d'instruction
  - Contraction d'une assurance
- CNIL
- Les données ne doivent pas être utilisées pour/contre le patient.

Un corpus difficile à constituer

Participation des patients :

- Identification des patients capables de participer
- Identifier ceux qui acceptent de participer
- Gérer les inquiétudes des patients
  - Identification des patients et obtention de leur accord (crainte de divulgation d'éléments biographiques) (perte $\sim$ 45 %)
  - Abandons (perte $\sim$ 10 %)

Un corpus difficile à constituer

- Transcription
  - Automatique : tests de plusieurs systèmes entraînés pour le français oral non concluants
  - Manuelle : deux annotateurs (dont le/la psychologue)
- Guide de transcription

Plan

Anonymisation traditionnelle

- Identification et substitution des entités nommées
  - outil performant mais non opérationnel [Grouin and Zweigenbaum, 2013]

  → scripts Python

- Intervention humaine importante
  - 10 catégories
  - identification par les majuscules
  - vérification manuelle

spk1 Donc vous habitez sur Ville1 ↑
spk2 Ville2
spk1 Ville2 D'accord alors moi qui suis pas du tout d'ici, c'est pas très loin

- Capacité d'ajouter des "bip" sonores sur la bande son

Normalisation des corpus

- Normalisation : trentaine d'expressions régulières
  sous-corpus Ville2 : `MS Word` → `Distagger`
  sous-corpus Ville1 : `CLAN` → `Distagger`
- Identification des tours de parole par numéro unique
- Statut du sujet (patient *vs* témoin) : dans la structure du corpus
- Psychologue pris en compte pour la dynamique de l'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 Ville1 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 Ville2**

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 Ville2** 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 Ville2**

Impossibilité de l'anonymisation

- Tâche avec faible contexte : *randomiser* les tours de paroles
- Impossibilité d'anonymiser l'histoire et la géographie
  - limiter le nombre d'intervenants sur la ressource : analyses sémantico-pragmatiques, bandes sons, etc.

De la réalité des patients

- Analyse formelle de le langage = définir une norme
- Dévier = dysfonctionnement

- Or, tout locuteur est confronté quotidiennement à des troubles du langage provenant de personnes saines.

- Le diagnostic ne peut souffrir d'approximations
  (comme celles des outils du TAL).

Conclusion

Difficultés :

- Obtenir les autorisations
- Accès aux patients et leur gestion
- Respecter l'anonymat des personnes sans renoncer aux enjeux scientifiques
- Identification des conséquences des conclusions


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


Chaika, E. (1974).
A linguist looks at “schizophrenic” language.
Brain and Language, 1(3) :257–276.
4


Fromkin, V. A. (1975).
A linguist looks at “a linguist looks at ‘schizophrenic language’”.
Brain and Language, 2(0) :498 – 503.
4


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.
19


Musiol, M. and Trognon, A. (1996).
L'accomplissement interactionnel du trouble schizophrénique.
Raisons Pratiques 7, pages 179–209.
4


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.

4