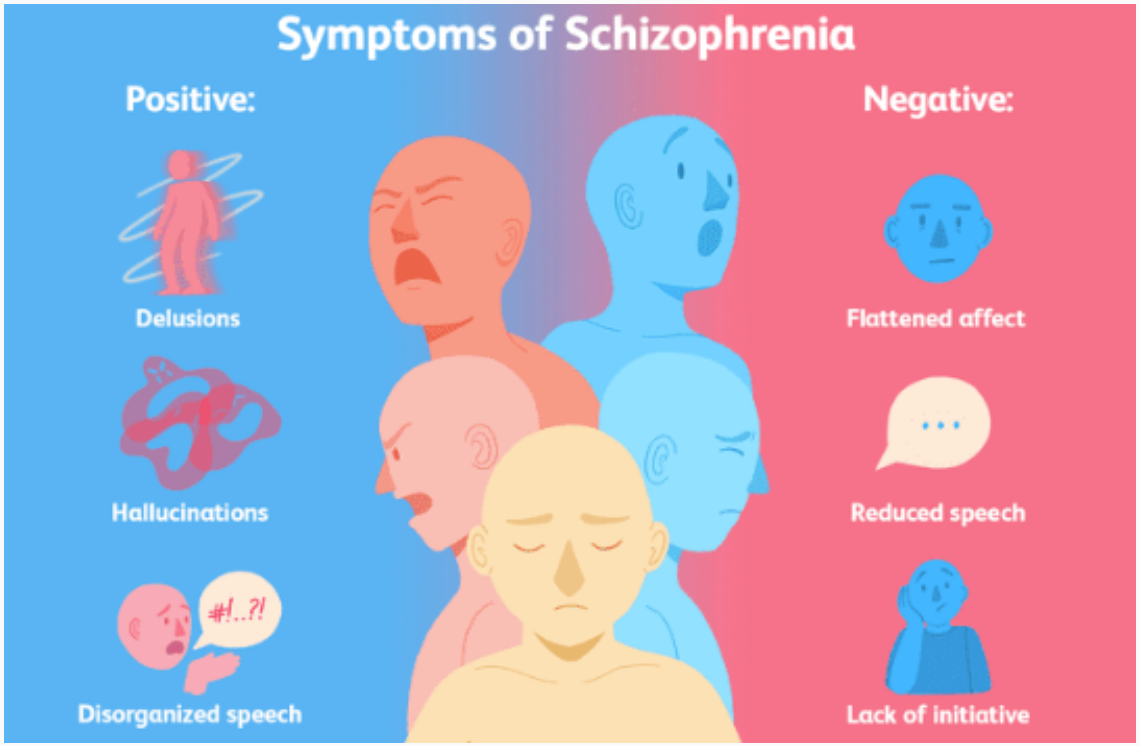
Investigation des marqueurs langagiers non-lexicaux et spécifiques des personnes souffrant de schizophrénie dans des conversations spontanées

Chuyuan Li[1] Maxime Amblard[1] Chloé Braud[2]
Caroline Demily[3] Nicolas Franck[3] Michel Musiol[1,4]

(1) Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
(2) éq. MELODI, IRIT, Université de Toulouse, CNRS, Toulouse, France
(3) Centre Hospitalier le Vinatier & UMR 5229, CNRS - Univeristé Lyon 1, Lyon, France
(4) Université de Lorraine, CNRS, ATILF, UMR 7118, F-54000 Nancy, France

Schizophrénie

- Trouble mental sévère
- Symptômes : idées délirantes, hallucinations, discours désorganisé



Source: *https://www.verywellmind.com/what-are-the-symptoms-of-schizophrenia-2953120*

Schizophrénie

- Enjeu
  - aide au diagnostic des médecins
  - amélioration de la compréhension du fonctionnement du langage en général
  - adaptation des systèmes de TAL à des parties de la population susceptible de développer la maladie

Description des données : Projet SLAM

- **S**chizophrénie et **L**angage : **A**nalyse et **M**odélisation [Rebuschi et al., 2014, Amblard et al., 2015]
- Entretiens semi-dirigés entre 1 psychologue (PSY) et
  - 18 SCZ
  - 23 témoins (TEM) : étudiants, biais lexicaux
  - dans chaque groupe 15 hommes, biais de genre
- Enregistrés avec un double système d'*eye-tracker* (données non-utilisées ici)
- Thématique abordée : le quotidien du participant

État de l'art

Classification automatique de SCZ[1] fondée sur des données langagières :

	type de données	langue	traits	res.
[Strous et al., 2009]	écrits	en	lexicaux	Acc. = 83,3%
[Mitchell et al., 2015]	*tweets*	en	lexicaux	Acc. = 82,3%
[Kayi et al., 2017]	écrits et *tweets*	en	morpho-synt. syntaxiques	F1 = 81,65%
[Allende-Cid et al., 2019]	textes narratifs	en	morpho-synt.	F1 = 82,8%
[Amblard et al., 2020]	conversations cliniques	fr	lexicaux	Acc. = 93,7%

[1]SCZ : personnes avec schizophrénie

État de l'art

Classification automatique de SCZ[1] fondée sur des données langagières :

	type de données	langue	traits	res.
[Strous et al., 2009]	écrits	en	lexicaux	Acc. = 83,3%
[Mitchell et al., 2015]	*tweets*	en	lexicaux	Acc. = 82,3%
[Kayi et al., 2017]	écrits et *tweets*	en	morpho-synt. syntaxiques	F1 = 81,65%
[Allende-Cid et al., 2019]	textes narratifs	en	morpho-synt.	F1 = 82,8%
[Amblard et al., 2020]	conversations cliniques	fr	lexicaux	Acc. = 93,7%

$\Rightarrow$ Corpus de nature différente : comparaisons difficiles

[1] SCZ : personnes avec schizophrénie

Exemples de données

PSY-SCZ

PSY : Et donc là vous allez voir un atelier euh... c'est quoi c'est...

SCZ : Oui donc là je suis allé en atelier thérapeutique euh euhh comment ils appellent ça... pas entretien thérapeutique... j'ai euh...

PSY : Education thérapeutique... c'est ça

PSY-TEM

PSY : Vous voulez faire quoi après

TEM : Euhh je voudrais faire le master de N. de psychopatho de la cognition et des interactions

PSY : Mmh mmh

Table de matières

Approches

Représentation des données

- Focaliser sur les *tours de parole* (TDP) des patients
- Enjeu : rareté des données $\Rightarrow$ 41 doc., 115k mots, 10k TDP

Représentation des données

- Focaliser sur les *tours de parole* (TDP) des patients
- Enjeu : rareté des données $\Rightarrow$ 41 doc., 115k mots, 10k TDP
- Plus contrôle sur la dispersion des données
  1. Mode "`Indiv.`" : TDP individuels
  2. Mode "`Full`" : concaténation des TDP

	#Doc.	#TDP / doc.			#Mot / doc.		
Config.	total	min	max	moy.	min	max	moy.
Indiv.	10,319	1	1	1	1	274	11
Full	41	76	555	268	703	6,778	2,811

Représentation des données

- Focaliser sur les *tours de parole* (TDP) des patients
- Enjeu : rareté des données $\Rightarrow$ 41 doc., 115k mots, 10k TDP
- Plus contrôle sur la dispersion des données
  1. Mode "`Indiv.`" : TDP individuels
  2. Mode "`Full`" : concaténation des TDP
  3. Mode "`W-n`" ($n \in \{128, 256, 512\}$) : TDP de taille au max. n tokens

	#Doc.	#TDP / doc.			#Mot / doc.		
Config.	total	min	max	moy.	min	max	moy.
Indiv.	10,319	1	1	1	1	274	11
W-128	893	1	34	11	128	317	145
W-256	443	1	72	20	256	424	271
W-512	209	2	129	42	512	609	530
Full	41	76	555	268	703	6,778	2,811

Modélisation : traits utilisés

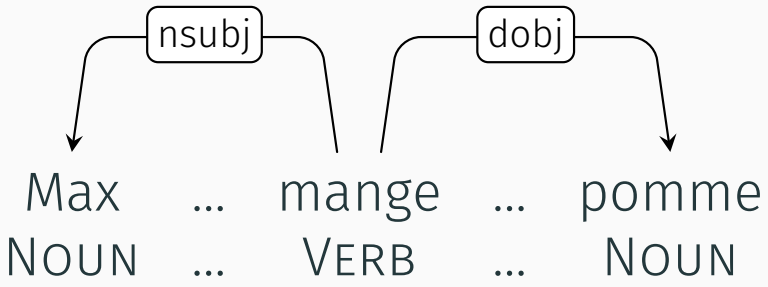
- Traits dialogiques
  - *Open Class Repair* (OCR) [Howes et al., 2012] : “pardon ?”, “huh ?”, “vous disiez ?”, *etc.*
  - *Backchannel réponse* (BC) : “ouais”, “hum mmh”, *etc.*
  - Connecteurs dans *LexConn* [Roze et al., 2012] : “parce que”, “mais”, “depuis que”, *etc.*

Modélisation : traits utilisés

- Traits dialogiques
  - *Open Class Repair* (OCR) [Howes et al., 2012] : “pardon ?”, “huh ?”, “vous disiez ?”, *etc.*
  - *Backchannel réponse* (BC) : “ouais”, “hum mmh”, *etc.*
  - Connecteurs dans *LexConn* [Roze et al., 2012] : “parce que”, “mais”, “depuis que”, *etc.*
- Traits délexicalisés / morpho-syntaxiques, UDPipe [Straka and Straková, 2017]
  - n-gramme *Part-of-speech* (POS tag), $n \in \{1, 2, 3\}$
  - *treelet* ($n \in \{2, 3\}$), [Johannsen et al., 2015]

Modélisation : traits utilisés

- Traits dialogiques
  - *Open Class Repair* (OCR) [Howes et al., 2012] : “pardon ?”, “huh ?”, “vous disiez ?”, *etc.*
  - *Backchannel réponse* (BC) : “ouais”, “hum mmh”, *etc.*
  - Connecteurs dans *LexConn* [Roze et al., 2012] : “parce que”, “mais”, “depuis que”, *etc.*
- Traits délexicalisés / morpho-syntaxiques, UDPipe [Straka and Straková, 2017]
  - *n*-gramme *Part-of-speech* (POS tag), $n \in \{1, 2, 3\}$
  - *treelet* ($n \in \{2, 3\}$), [Johannsen et al., 2015]
  - un exemple :



NOUN, NOUN-VERB,
NOUN-VERB-NOUN

VERB $\xrightarrow{\text{Nsubj}}$ NOUN

NOUN $\xleftarrow{\text{Nsubj}}$ VERB $\xrightarrow{\text{Dobj}}$ NOUN

Expériences

Classification

- Problème : peu de données, dimensions très élevées
  - Sélection de traits avec *feature_selection.SelectFromModel*[2]
  - Validation croisée enchâssée pour diviser train/test

[2] *https://scikit-learn.org/*

Classification

- Problème : peu de données, dimensions très élevées
  - Sélection de traits avec `feature_selection.SelectFromModel`[2]
  - Validation croisée enchâssée pour diviser train/test

- 5 Classifieurs scikit learn
  - *Naive Bayes*
  - Régression logistique
  - SVM
  - *Random Forest*
  - Perceptron

[2] `https://scikit-learn.org/`

Résultats

Différents jeux de traits

Exactitude moyenne pour la configuration *Full*, *Indiv.* et *W-n* (*bow* et *ngram* viennent de [Amblard et al., 2020]) :

Traits	Full	Indiv.	W-128	W-256	W-512
bow	93.66	72.43	-	-	-
ngram	85.61	69.59	-	-	-
OCR	60.62	50.17	52.43	55.19	59.28
BC	74.48	54.79	62.01	66.89	67.86
Connectives	72.44	55.28	64.05	69.68	73.57
POS	53.66	55.80	60.63	60.48	60.09
2-POS	67.36	56.33	64.85	68.53	71.74
3-POS	71.65	56.53	65.39	70.66	72.55
2-treelet	69.19	56.73	65.02	70.11	74.19
3-treelet	66.78	55.34	63.95	66.39	69.03
1-2-3-POS	69.01	58.36	66.19	72.03	72.67
POS+2-3-treelet	66.59	57.77	65.52	69.11	72.39
3-POS+BC	74.93	57.46	69.92	73.75	77.86

Comparaison

	type de données	taille	traits	res.
[Mitchell et al., 2015]	*tweets*	$1,1m$ *tweets*[3]	LIWC+LDA	Acc. = 82,3%
[Howes et al., 2012]	conversations	131 doc[4]	bow	Acc. = 93,0%[5]
nous	conversations	41 doc[6]	bow	Acc. = 93,66%
[Kayi et al., 2017]	rédaction	373 doc	POS	F1 = 69,76%
	tweets	$974k$ *tweets*	POS	F1 = 69,20%
[Allende-Cid et al., 2019]	textes narratifs	189 textes	meta-POS	F1 = 75,1%
nous	conversations	41 doc	3-POS	F1 = 74,34%

[3] limitation à 140 caractères/*tweet*
[4] moy. 320 TDP/doc
[5] [Howes et al., 2012] classifient l'adhérence au traitement
[6] moy. 268 TDP/doc, exclure PSY.

Traits lexicaux [Amblard et al., 2020]

Features	Full	Indiv.	W-*n*
bow	93.66	72.43	-
ngram	85.61	69.59	-

- *bow* et *ngram*, très bons indicateurs

Traits lexicaux [Amblard et al., 2020]

Features	Full	Indiv.	W-*n*
bow	93.66	72.43	-
ngram	85.61	69.59	-

- *bow* et *ngram*, très bons indicateurs
  - SCZ, équivalent à [Strous et al., 2009, Mitchell et al., 2015]
    - thèmes maladie : “douleur”, “hospitalisé”, “hallucinations”, *etc.*
    - + déictiques à la 1^e pers. : “je / j’”, “mon”
  - TEM
    - thèmes éducation et psychologie : “fac”, “master”, “psychologue”, *etc.*
    - + déictiques à la 2^e pers. : “tu / t’”, “vous”
- dû à la nature des données, mais... à généraliser ?

Traits syntaxiques

Traits	Full	Indiv.	W-128	W-256	W-512
POS	53.66	55.80	60.63	60.48	60.09
2-POS	67.36	56.33	64.85	68.53	71.74
3-POS	71.65	56.53	65.39	70.66	72.55
2-treelet	69.19	56.73	65.02	70.11	74.19
3-treelet	66.78	55.34	63.95	66.39	69.03
1-2-3-POS	69.01	58.36	66.19	72.03	72.67
POS+2-3-treelet	66.59	57.77	65.52	69.11	72.39
3-POS+BC	74.93	57.46	69.92	73.75	77.86

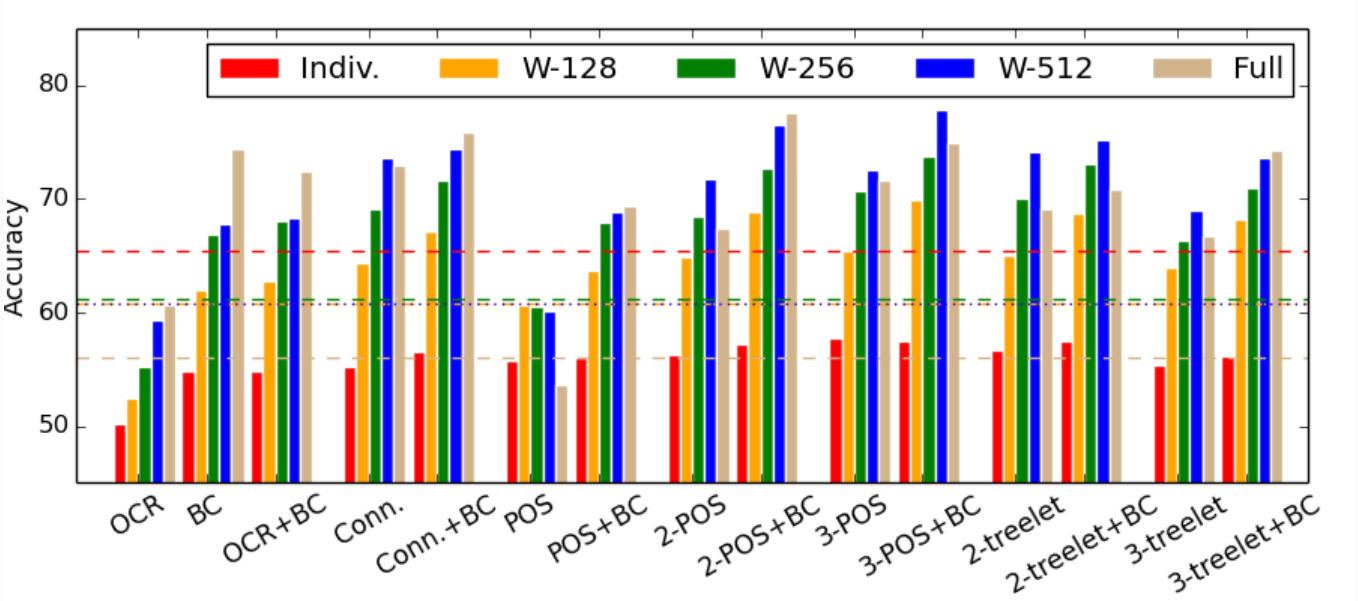
- *POS* et *treelet*, bons indicateurs

Traits syntaxiques

Traits	Full	Indiv.	W-128	W-256	W-512
POS	53.66	55.80	60.63	60.48	60.09
2-POS	67.36	56.33	64.85	68.53	71.74
3-POS	71.65	56.53	65.39	70.66	72.55
2-treelet	69.19	56.73	65.02	70.11	74.19
3-treelet	66.78	55.34	63.95	66.39	69.03
1-2-3-POS	69.01	58.36	66.19	72.03	72.67
POS+2-3-treelet	66.59	57.77	65.52	69.11	72.39
3-POS+BC	74.93	57.46	69.92	73.75	77.86

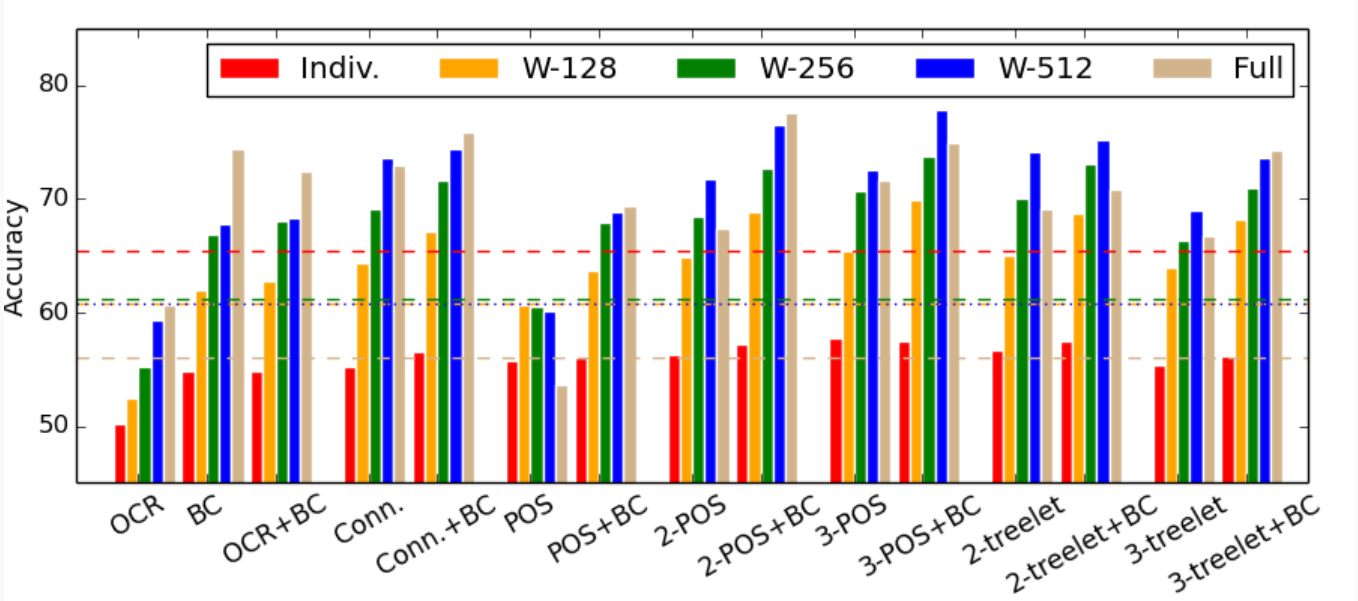
- *POS* et *treelet*, bons indicateurs
  - SCZ : + structure VERBALE (aussi dans [Kayi et al., 2017])
    - VERB $\xrightarrow{\text{Aux}}$ AUX (Ex. : "(j')ai fait", "(c')est (pas) gagné")
    - VERB $\xrightarrow{\text{Nsubj}}$ PRON (Ex. : "ça va", "(je) sais pas")
  - TEM : + structure compliquée
    - "SCONJ" : conjonction subordonnée
    - "CCONJ" : conjonction coordonnée

Dialogue et discours



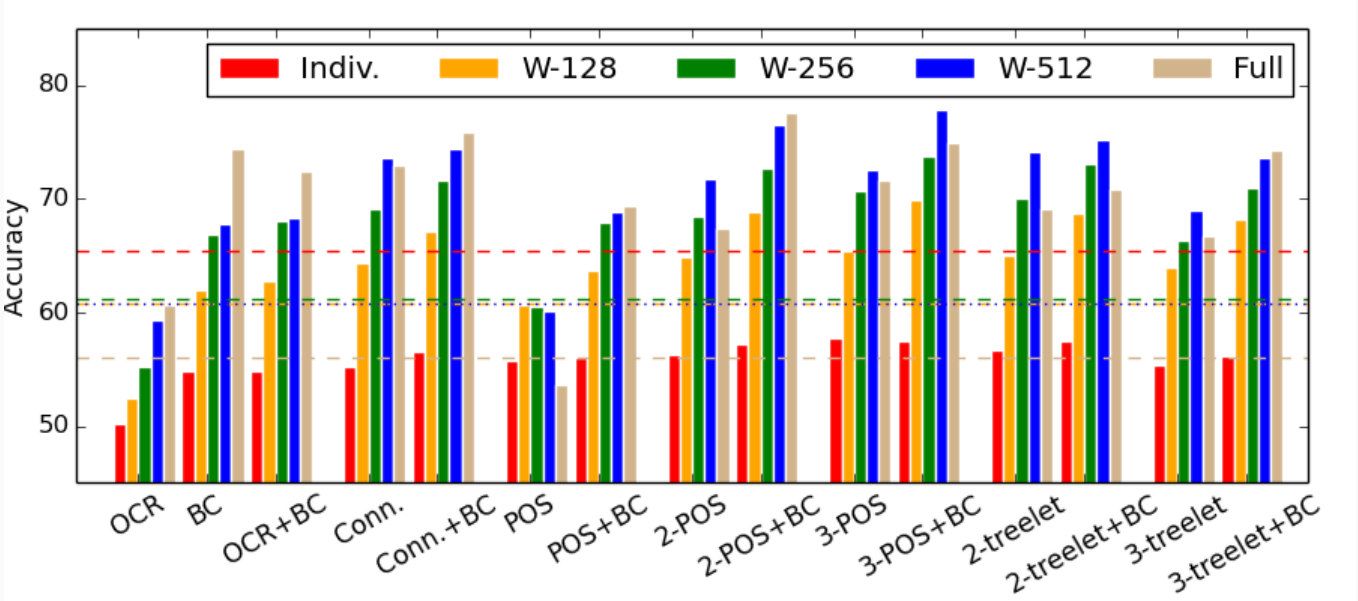
- *OCR*, mauvais résultats

Dialogue et discours



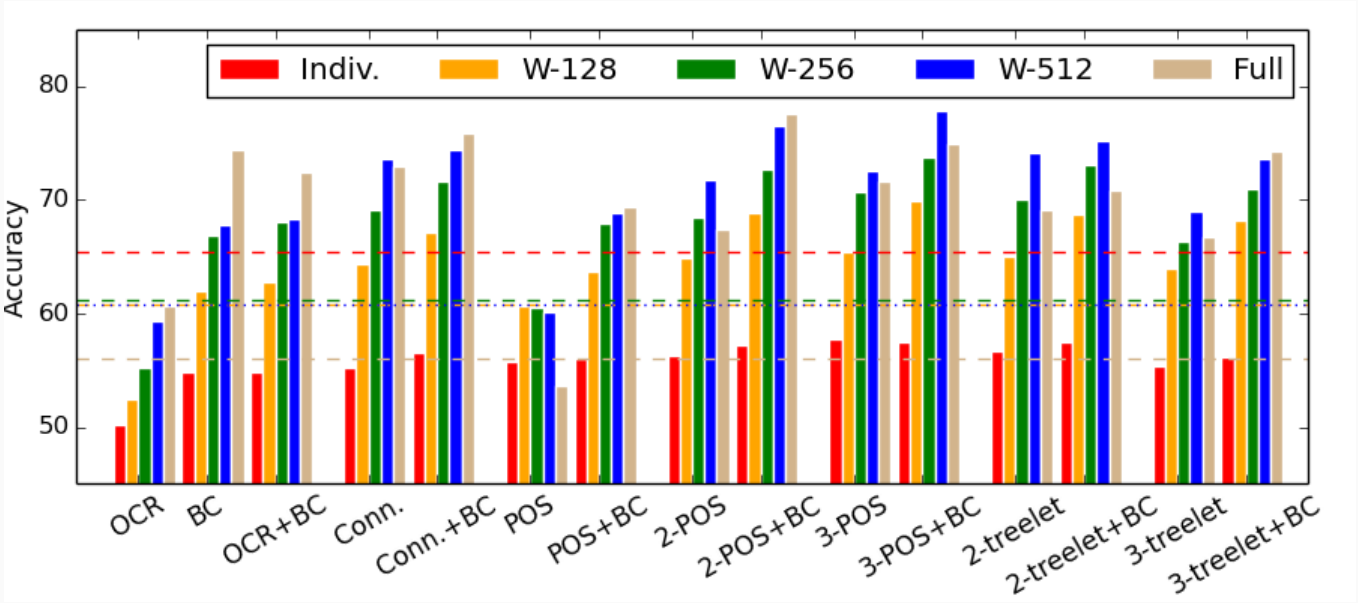
- *OCR*, mauvais résultats
- *Backchannel*, améliore le résultat systématiquement (voir les combinaisons)
  - TEM : “ah, hum-hum”, phatiques
  - SCZ : “je comprends, exactement”, flou

Dialogue et discours



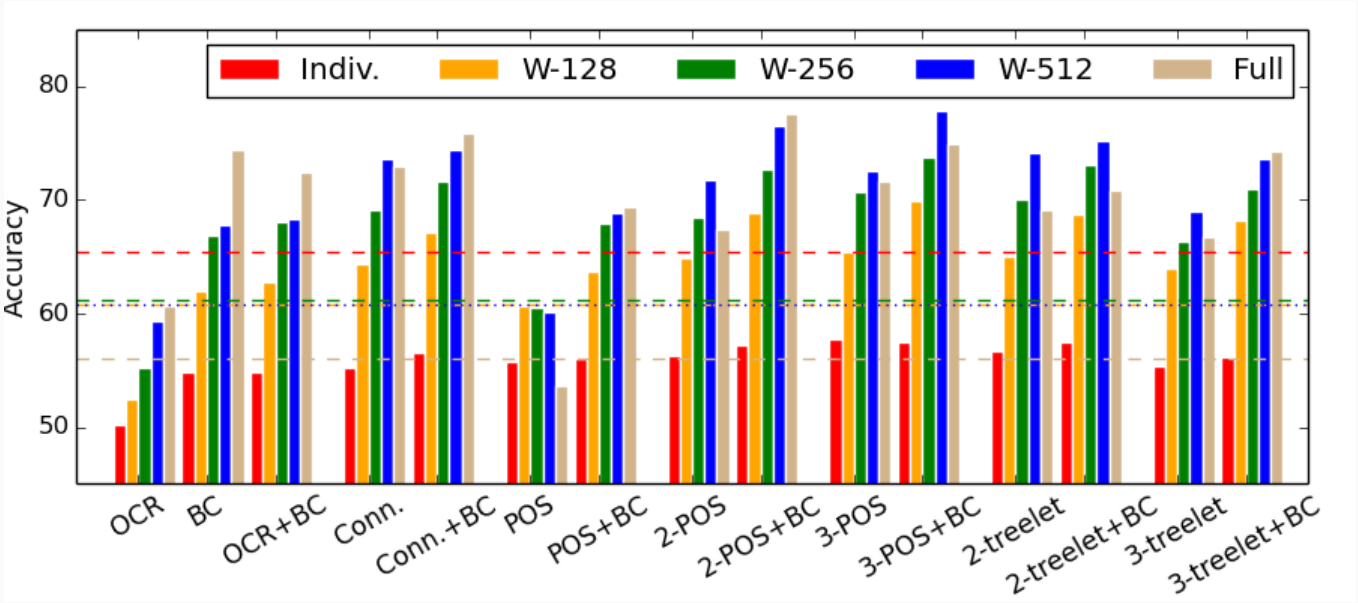
- *OCR*, mauvais résultats
- *Backchannel*, améliore le résultat systématiquement (voir les combinaisons)
  - TEM : “ah, hum-hum”, phatiques
  - SCZ : “je comprends, exactement”, flou
- *Connecteurs*, bon indicateur
  - TEM : “jusqu’à ce que, au point de”, +longs
  - SCZ : “maintenant, depuis”, présent

Taille du contexte



- En général contexte +long, +score : Indiv. < W128 < W256 < W512

Taille du contexte



- En général contexte +long, +score : Indiv. < W128 < W256 < W512
- Mais W512 > Full $\Rightarrow$ dispersion des données

Conclusion

Conclusion

- Premier système en français
- Test de différentes représentations du contexte et de traits linguistiques
- Biais lexicaux dans les deux groupes
  $\Rightarrow$ Exploration de traits “haut niveau” (moins dépendants de la langue)

Conclusion

- Premier système en français
- Test de différentes représentations du contexte et de traits linguistiques
- Biais lexicaux dans les deux groupes
  ⇒ Exploration de traits “haut niveau” (moins dépendants de la langue)
- Perspectives
  - Amélioration des traits dialogiques via la désambiguïsation
  - Contexte complexe : ajout de l’interaction

Merci !

Liste OCR et BC

Ces listes ont été obtenues en traduisant celles fournies par les auteurs de [Howes et al., 2012].

Table 1: *Open Class Repair*

pardon vous disiez	pardon
ah vous parler pardon	excusez-moi
excuse moi	bon je suis désolée
désolé(e)	(ah) ouais ?
ah bon ?	c'est vrai ?
c'est euh ?	hum ?
de quoi	c'est quoi ?
c'est-à-dire	euh ?
dites moi plus	mais encore

Table 2: *Backchannel*

oui	ouais	ouais voilà
oui c'est ça	oui bah oui	oui… forcément
bah ouais	hum (hum)	muh mmh
mmh/mmhh	d'accord	ok
voilà	c'est ça	c'est vrai
c'est sûr	ça c'est clair	eh bien sûr
carrément	bien sûr	super
ok… bon	d'accord ça marche	certes
mais hein	je comprends	vraiment
bien	bon	très bien
quand même	tout à fait	certainement
exactement	tant mieux	oh
ah	ben	alors ben
ah d'accord	ah ça euh	eh bah c'est bien

Hyper-paramètres

Pendant l'apprentissage :

- *Naive Bayes*: lissage
  $\alpha \in V = \{0.001, 0.005, 0.01, 0.1, 0.5, 1, 5, 10, 100\}$;
- Régression logistique : norme L_2 et optimisation de $C \in V$;
- SVM avec kernel linéaire : norme L_2 et optimisation de
  $C \in V \cup \{1000\}$;
- *Random Forest*: max_depth $\in \{2, None\}$;
- Perceptron: norme L_2 et lissage $\alpha \in V$;

Références i

Allende-Cid, H., Zamora, J., Alfaron-Faccio, P., and Alonso, M. (2019).
A machine learning approach for the automatic classification of schizophrenic discourse.
IEEE Access, pages 45544–45554.

Amblard, M., Braud, C., Li, C., Demily, C., Franck, N., and Musiol, M. (2020).
Investigation par méthodes d'apprentissage des spécificités langagières propres aux personnes avec schizophrénie (investigating learning methods applied to language specificity of persons with schizophrenia).
In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants*

Références ii

Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2: Traitement Automatique des Langues Naturelles, pages 12–26.

Amblard, M., Fort, K., Demily, C., Franck, N., and Musiol, M. (2015). **Analyse lexicale outillée de la parole transcrite de patients schizophrènes.** *Traitement Automatique des Langues*, 55(3):91 – 115.

Howes, C., Purver, M., McCabe, R., Healey, P., and Lavelle, M. (2012). **Predicting adherence to treatment for schizophrenia from dialogue transcripts.** In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 79–83.

Références iii


Johannsen, A., Hovy, D., and Søgaard, A. (2015).
Cross-lingual syntactic variation over age and gender.
In *Proceedings of the nineteenth conference on computational natural language learning*, pages 103–112.


Kayi, E. S., Diab, M., Pauselli, L., Compton, M., and Coppersmith, G. (2017).
Predictive linguistic features of schizophrenia.
In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017)*, pages 241–250.


Mitchell, M., Hollingshead, K., and Coppersmith, G. (2015).
Quantifying the language of schizophrenia in social media.
In *Proceedings of the 2nd workshop on Computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 11–20.

Références iv

Rebuschi, M., Amblard, M., and Musiol, M. (2014).
Using SDRT to analyze pathological conversations. Logicality, rationality and pragmatic deviances.
In Rebuschi, M., Batt, M., Heinzmann, G., Lihoreau, F., Musiol, M., and Trognon, A., editors, *Interdisciplinary Works in Logic, Epistemology, Psychology and Linguistics: Dialogue, Rationality, and Formalism*, volume 3 of *Logic, Argumentation & Reasoning*, pages 343 – 368. Springer.

Roze, C., Danlos, L., and Muller, P. (2012).
Lexconn: a french lexicon of discourse connectives.
Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics, (10).

Références v

Straka, M. and Straková, J. (2017).
Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe.
In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Strous, R. D., Koppel, M., Fine, J., Nachliel, S., Shaked, G., and Zivotofsky, A. Z. (2009).
Automated characterization and identification of schizophrenia in writing.
The Journal of nervous and mental disease, 197(8):585–588.