Cognition & Linguistique Computationnelle

Olga Seminck

Université catholique de Louvain
Institute of Neuroscience
Media innovation and intelligibility Lab
Centre de traitement automatique du langage

olga.seminck@uclouvain.be

Louvain-la-Neuve, 31 Octobre 2019

Plan

Introduction

Le langage est une faculté cognitive

la mémoire
le langage
le raisonnement
l'apprentissage
l'intelligence
la résolution de problème
la prise de décision
la perception ou l'attention

`https://fr.wikipedia.org/wiki/Cognition`

Questions linguistiques-cognitives

Questions de **mémoire** :
Quel est rôle de la mémoire dans la compréhension et la production du langage ?

Question de **raisonnement** :
Est-ce que le langage est efficace comme système de communication ?

Question d'**intelligence** :
Comment l'intelligence est-elle reflétée à travers du langage ?

Question sur la **complexité** par rapport aux capacités cognitives :
Comment nos capacités cognitives déterminent-elles le système du langage ?

Accès à la faculté du langage

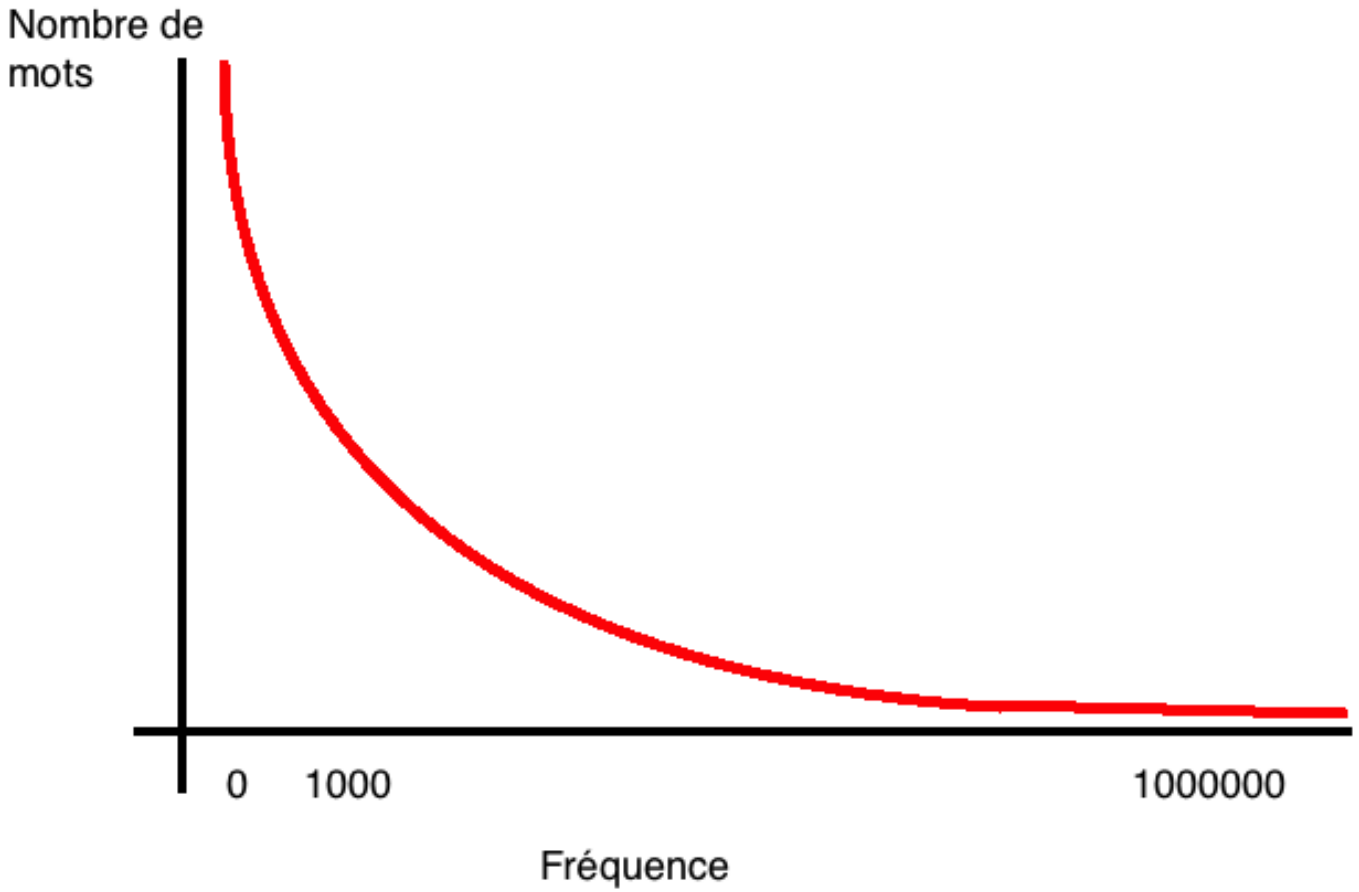
Les productions de langage

Les réactions face aux stimuli langagiers

- temps de lecture
- temps de réaction (appuyer sur un bouton)
- dilatation des pupilles
- Event-related potentials (ERPs)
- ...

Le langage possède des propriétés probabilistes

Loi de Zipf



Le langage, en tant que faculté cognitive, est sensible aux fréquences

Métrique de coût: formule qui prédit le coût cognitif

- Traduit une hypothèse dans une prédiction

Exemple: la surprise

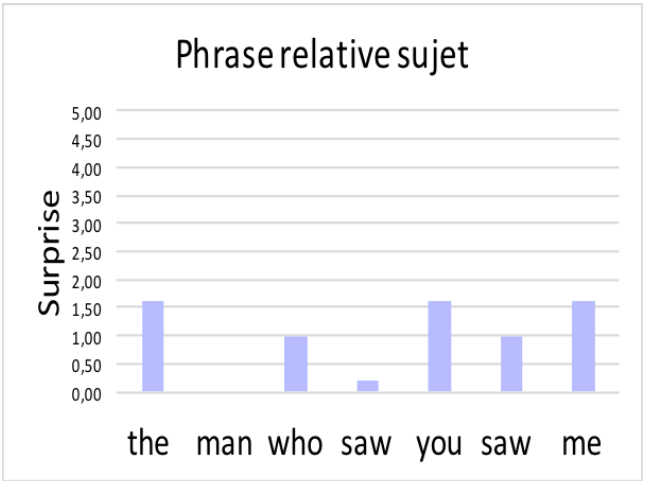
- Hypothèse: les événements non-attendus sont plus difficile à traiter
- Métrique de coût:
  Difficulté(event) $= -\log(P(event))$

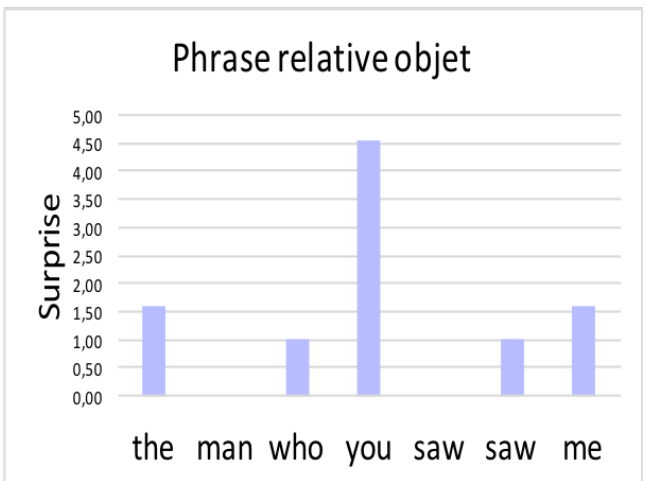
La théorie de l'information (Shannon and Weaver 1949) est l'inspiration pour des métriques de coût pour des processus linguistiques

La Théorie de la Surprise

La surprise simule la difficulté de traitement des phrases relatives sujet et objet. (Hale 2001)

$Surprisal(event) = -log(P(event))$





Règle			probabilité
NP	→	SPECNP NBAR	0, 33
NP	→	*you*	0, 33
NP	→	*me*	0, 33
SPECNP	→	DT	1, 0
NBAR	→	NBAR S[+R]	0, 5
NBAR	→	N	0, 5
S	→	NP VP	1, 0
S[R+]	→	NP[+R] VP	0, 869
S[R+]	→	NP[+R] S/NP	0, 131
S/NP	→	NP VP/NP	1, 0
VP/NP	→	V NP/NP	1, 0
VP	→	V NP	1, 0
V	→	*saw*	1, 0
NP[+R]	→	*who*	1, 0
DT	→	*the*	1, 0
N	→	*man*	1, 0
NP/NP	→	ϵ	1, 0

Linguistique computationnelle pour la recherche cognitive

La sensibilité de la faculté du langage aux propriétés statistiques

$\rightarrow$

Beaucoup de possibilités pour l'étude du langage en tant que faculté cognitive par des méthodes computationnellles

Le TAL : expertise en analyse computationnelle

- Applicable à des grands corpus
- Robustesse

Deux études Cognition & Linguistique Computationnelle

- Un métrique de coût cognitif de la résolution des pronoms
- La détection automatique de la maladie d'Alzheimer dans les conversations électroniques

Les modèles cognitifs et computationnels de la résolution des pronoms

La résolution d'anaphores

La résolution des pronoms est une forme de résolution anaphorique.

SN α_1 prend SN α_2 comme son antécédent anaphorique si α_1 dépend de α_2 pour son interprétation.
(Van Deemter and Kibble 2000)

A **secret**'s worth depends on the people from whom **it** must be kept.

The Shadow of the Wind, Carlos Ruiz Zafón

La résolution des pronoms est le processus qui consiste en trouver l'antécédent d'un pronom anaphorique.

Expérience:

1. Un métrique de coût cognitif pour la résolution des pronoms
&
2. Des preuves des données oculométriques

1. Un métrique de coût cognitif pour la résolution des pronoms

Les métriques de coût

Métrique de coût: formule qui prédit le coût cognitif

- Traduit une hypothèse dans une prédiction

Exemple: la surprise

- Hypothèse: les événements non-attendus sont plus difficile à traiter
- Métrique de coût :
  Difficulté(event) $= -\log(P(event))$

La théorie de l'information (Shannon and Weaver 1949) est l'inspiration pour des métriques de coût pour des processus linguistiques

Métrique de coût pour la résolution des pronoms

- Basé sur l'entropie

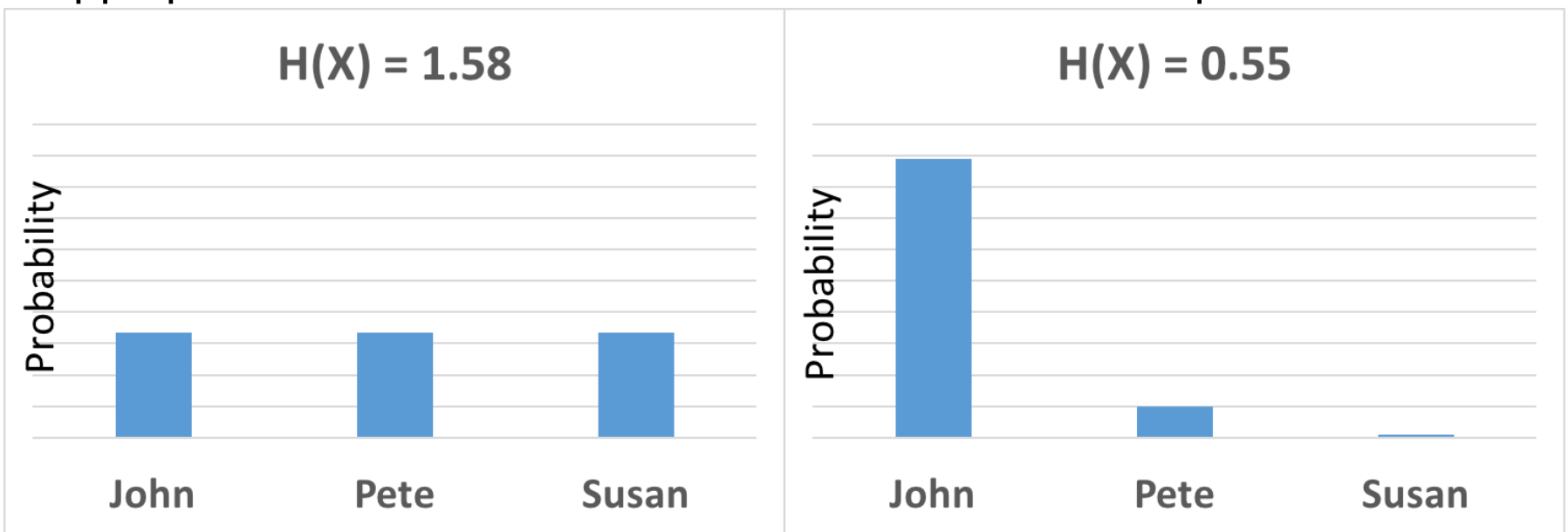
Un métrique de coût pour prédire la difficulté des pronoms

Prédiction pour la résolution des pronoms:
Plus d'incertitude sur l'antécédent $\rightarrow$ plus de coût cognitif

Entropie : mesure d'incertitude
$H(X) = -\sum_{j \in X} p(X = j) \cdot \log_2(p(X = j))$

Entropie

S'applique à une variable aléatoire : l'antécédent d'un pronom



L'entropie relative

L'entropie augmente quand le nombre d'antécédents potentiels augmente.

- Il faut pouvoir comparer les scores dans un texte
- 'Normaliser' l'entropie

Normalisation: entropie relative
'Distance' entre la distribution de probabilités actuelle & une distribution plate

$$H_{relative}(P||Q) = \sum_{i \in P \wedge i \in Q} P(i) \log \frac{P(i)}{Q(i)} \quad (1)$$

Plus de distance $\Rightarrow$ moins d'incertitude $\Rightarrow$ moins de coût cognitif

Un système TAL donne la distribution des probabilités

The Red House tells the story of **a mysterious, tormented individual** who breaks into **toy shops and museums** to steal **dolls and puppets**. Once they are in his power...

1. Distribution de proba. depuis les paramètres du système
2. Calcul de l'entropie relative sur cette dist. proba.

Antécédent de they	Probabilité	Entropie relative
The Red House	0.05	
a mysterious, tormented individual	0.04	
toy shops and museums	0.31	0.83
dolls and puppets	0.58	
$\emptyset$	0.02	

Implémention

Obtenir des probabilités depuis un système TAL état de l'art (Lee et al. 2017)

Système de Lee et al.:

- Bout à bout : pas besoin de préprocessing
- Architecture réseau de neurone
- Système du type 'Ranking'

2. Des preuves des données oculométriques

Le Dundee Eye-Tracking Corpus (Kennedy et al. 2003)

Mouvements oculaires des 10 locuteurs natifs de l'anglais
Lecture de 65 textes
De 'the Independent' (journal)
Total: 50 000 tokens
Annoté avec des parties de discours (Frank 2010) et des relations syntaxiques en dépendance (Barrett et al. 2015)

Annotation de tous les 1 109 pronoms anaphoriques (Seminck and Amsili 2018)
Un jeu de données pour étudier la résolution des pronoms sur des données naturelles.

Lecture

Lecture: une séquence de fixations dans un texte.

Chaque fixation a une durée, exprimée en ms.

Les yeux ‘sautent’ de fixation en fixation.

(Rayner 1998)

216
Are tourists enticed by these attractions threatening their very existence ?

156
Are tourists enticed by these attractions threatening their very existence ?

227

Are tourists enticed by these attractions threatening their very existence ?

187

Are tourists enticed by these attractions threatening their very existence ?

182

Are tourists enticed by these attractions threatening their very existence ?

96

Are tourists enticed by these attractions threatening their very existence ?

232

Are tourists enticed by these attractions threatening their very existence ?

Are tourists enticed by these attractions threatening their very[335] existence ?

Are tourists enticed by these attractions threatening their[168] very existence ?

Are tourists enticed by these attractions threatening their very existence [173] ?

Are tourists enticed by these attractions threatening their very existence ?[188]

Are tourists enticed by these attractions threatening their very existence [88] ?

174
Are tourists enticed by these attractions threatening their very existence ?

168

Are tourists enticed by these attractions threatening their very existence ?

170

Are tourists enticed by these attractions threatening their very existence ?

271

Are tourists enticed by these attractions threatening their very existence ?

88

Are tourists enticed by these attractions threatening their very existence ?

232

Are tourists enticed by these attractions threatening their very existence ?

Are tourists enticed by these attractions threatening their very existence ?[202]

222

Are tourists enticed by these attractions threatening their very existence ?

Are tourists enticed by these attractions threatening their very existence [157] ?

Are tourists enticed by these attractions threatening their very existence [157] ?

Depuis la séquence de fixations, on peut mesurer le temps de lecture de plusieurs façons.

Temps de lecture : sommer la durée des fixation dans une région

Are | tourists | enticed | by | these | attractions | threatening | their | very | existence?

Souvent, on prend le mot comme région.

Présupposition : Plus de temps $\Rightarrow$ plus de difficulté de traitement

(Rayner 1998)

Exemple: *first pass & total reading time*

Are	tourists	enticed	by	these	attractions	threatening	their	very	existence?
1	2	3,13		4,14,15	5,16,17	6,7,18	9	8,19	10,11,12,20,21

First pass: $\sum$ durée des fixations 10, 11 et 12

Total: $\sum$ durée des fixations 10, 11, 12, 20 et 21

Mesurer le temps de lecture des pronoms : problématique

Seulement 20 - 30% des pronoms sont fixés.
(Ehrlich and Rayner 1983)

Les pronoms sont des mots courts.

Spill-over effects

Dans la littérature :
Prendre une fenêtre autour du pronom.

- ... at a time **[when they are at greatest risk]**, and then ...
- ... on it; **[but it would seriously degrade the]** quality ...

Soucis:

- Besoin de plusieurs modèles statistiques
- Moins de points de données par pronom

Solution: un métrique binaire

Est-ce que le pronom a été fixé ?

Variable bionomiale : réponse oui/non.

Advantages:

- Plus de données
- Seulement un endroit dans le texte où il faut mesurer

“ un mot est sauté, parce qu'il a été identifié lors de la fixation précédente.” (Brysbaert and Vitu 1998)

Hypothèse: un pronom fixé représente plus de difficulté de traitement.

Modèle Statistique

Modèle Statistique qui prédit si le pronom a été fixé.

Est-ce que l'entropie relative est pertinente dans cette prédiction ?

Modèle à effets mixte:
fixated $\sim$ length + frequency + comma + punctuation + rel_ent
+ (1 + rel_ent | participant) + (1 | dundee_tokens)

Résultat

Le métrique de coût basé sur l'entropie prédit la lecture

L'entropie relative était un prédicteur pertinent pour savoir si un pronom était fixé.

Une plus petite distance entre l'entropie et l'entropie maximale
$\Rightarrow$ plus de participants fixent le pronom

Estimation: -0.07 (Intervalle de crédibilité 95% = [-0.01, -0.13])

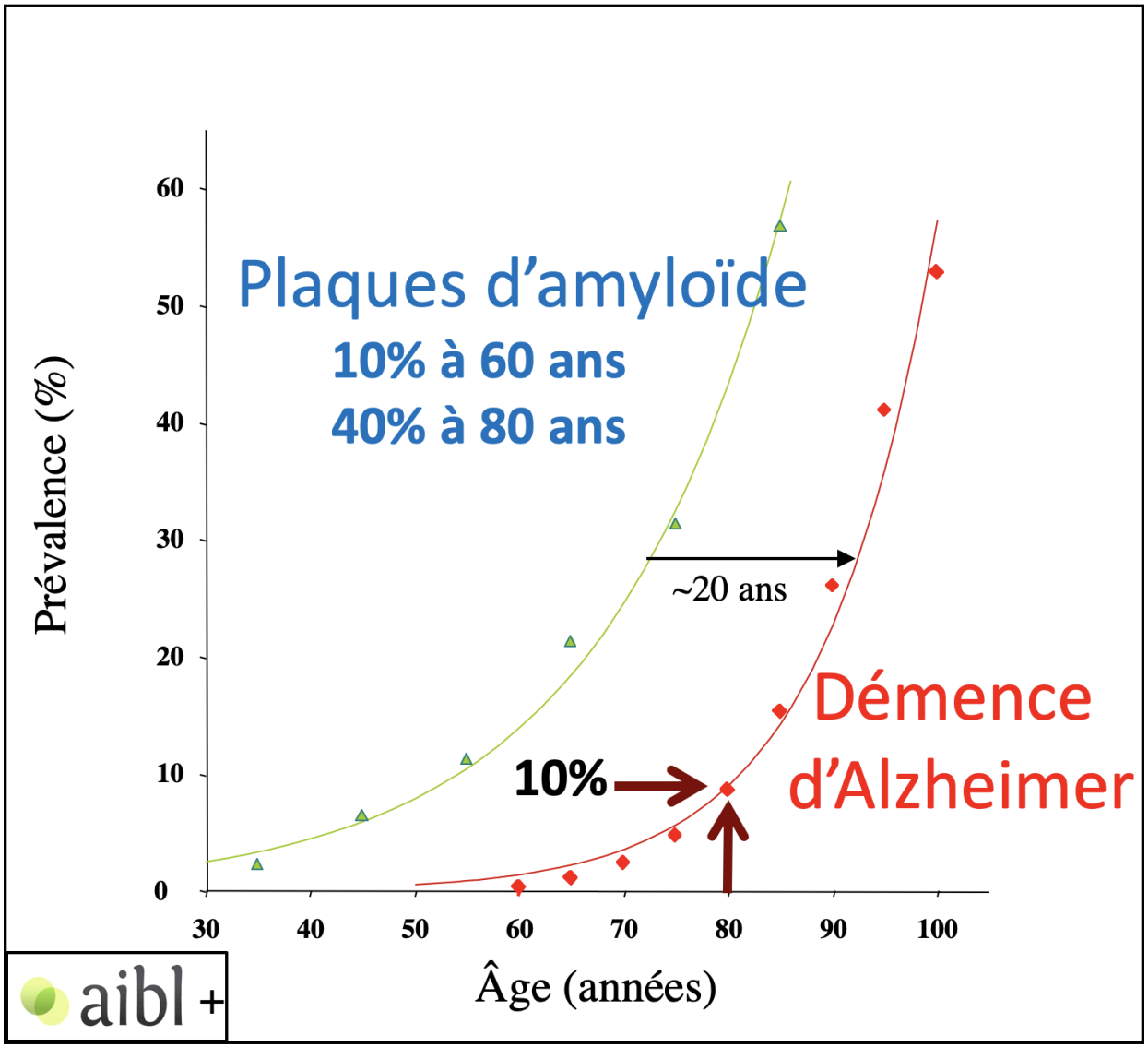
Conclusion:
La théorie de l'information est également pertinente dans la résolution des pronoms.

L'analyse automatique des messages électroniques pour la détection précoce de la maladie d'Alzheimer

La maladie d'Alzheimer

- La maladie d'Alzheimer est responsable de la majorité des cas de démence.(Swaab 2014)
- Aujourd'hui, il n'y a pas de médicaments efficaces sur le marché pour freiner l'évolution de la maladie.
- La maladie est marquée par l'apparence des plaques d'amyloïde et des agrégats de la protéine tau dans le cerveau.
- Les plaques d'amyloïde apparaissent 15 à 20 ans avant les symptômes. (McDade and Bateman 2017)

Apparition des plaques d'amyloïde (Rowe et al. 2010)



Adapté de Rowe et al *Neurobiology of Aging* (2010)

Une détection précoce est nécessaire

- Traiter les plaques chez les patients atteints de la démence s'est avéré inefficace. (McDade and Bateman 2017)
- Hypothèse : traiter les patients avant l'apparition des symptômes sera plus efficace.
- Moyens actuels de dépistage (Hanseeuw 2019) :
  - PET-scan pour détecter les plaques d'amyloïde (1500€ par scan)
  - Ponction lombaire (invasif pour le patient)
- 20% des personnes de 70 ans ont des plaques d'amyloïde (beaucoup d'argent gâché lors du recrutement)

Objectif du projet

Développer un outil qui contribue à la détection précoce qui soit :

- Bon marché
- Non-invasif
- Applicable à tout un chacun

Idée : utliser les historiques des messages électroniques pour la détection.

Biomarqueurs linguistiques de la maladie d'Alzheimer

Travaux sur la détection automatique de la maladie d'Alzheimer par le langage (en anglais) :

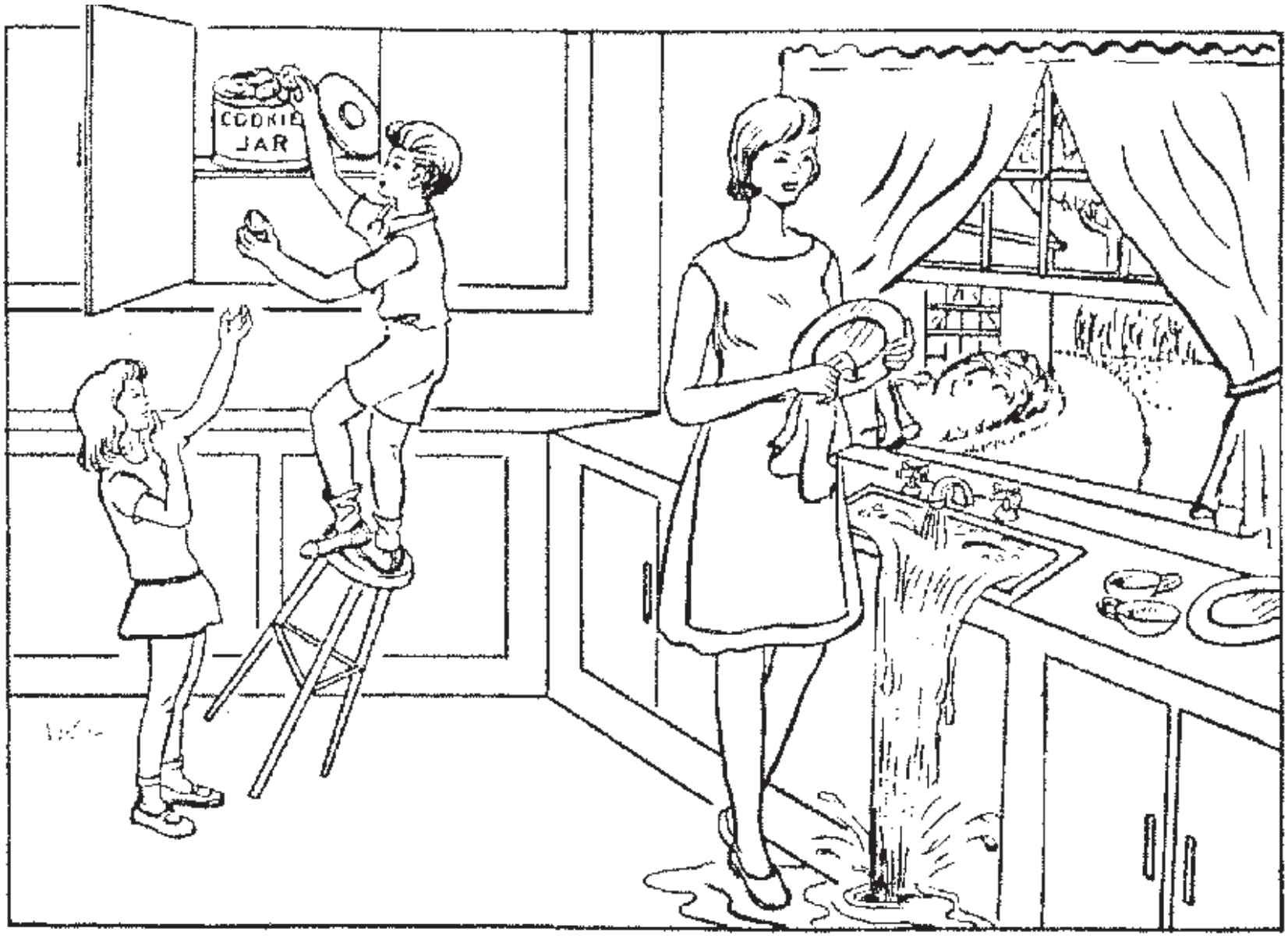
DementiaBank (MacWhinney et al. 2011; Becker et al. 1994) :

Cookie Theft Picture Task (Goodglass and Kaplan 1983) :
"Dites-moi tout ce qu'il se passe sur cette image."

240 descriptions de patients avec un diagnostic clinique d'Alzheimer

233 descriptions de participants sains

Cookie Theft Picture Task



Performances détection Cookie Theft Picture Task

Détection de la maladie d'Alzheimer dans les transcriptions des descriptions :

- 97.4 % de précision (accuracy) (Chen et al. 2019)

Un grand nombre de travaux en TAL sur cette ressource.

Plus de performance grâce aux architectures de réseaux de neurones.

Mais : à quel point ces modèles sont généralisable à d'autres types de corpus ?

Qu'est-ce que ça nous apprend sur la MA et la linguistique ?

Marqueurs linguistiques

Identification des marqueurs linguistiques de la maladie d'Alzheimer (Fraser et al. 2016)

- Perte de vocabulaire, usage de mots plus fréquents et usage de mots non existants
- Emploi des mots vides de sens truc, pronoms (il, elle, ça. . .)
- Répétition des mots et des phrases
- Structures syntaxiques moins riches
- Plus de phrases fragmentées
- Discours moins informatifs : moins d'éléments de l'image sont donnés

Exemple

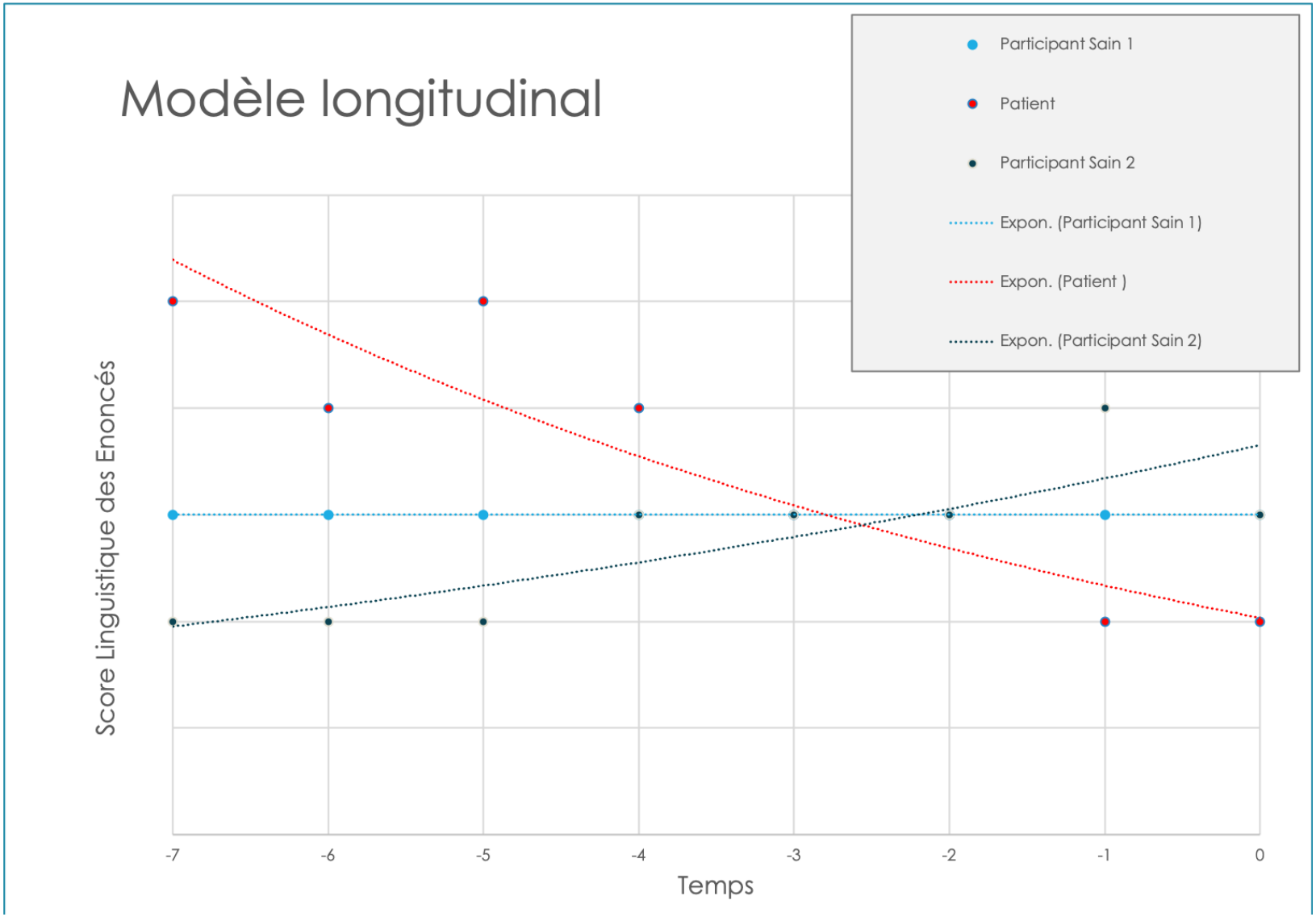
- *INV: okay what do you see going on in that picture ?
- *PAR: well this here is cookie jar . [+ gram]
- *INV: okay .
- *PAR: looks like the [/] the boy's pickin(g) up a bunch or something .
- *PAR: he's fallin(g) off the chair [: stool] [* s:r] down here or tryin(g) to .
- *INV: okay .
- *PAR: +< here and down hiss [* s:uk] . [+ jar]
- *PAR: she been washin(g) the dishes . [+ gram]

Nouveauté de notre projet

Utiliser l'historique des messages électroniques
(mail, WhatsApp, Messenger ...)

- Aspect longitudinal : comparaison d'une personne à elle-même
- Matériau déjà existant
- Analyse sur le français

Modèle Longitudinal



Méthode Phase 1 : Collection des données

- Copier les historiques des conversations électroniques :
  - 30 patients dans un stade précoce de la maladie d'Alzheimer (Cliniques de Saint Luc)
  - 30 participants sains (Université des Aînés)

Méthode Phase 2 : Constitution des corpus

- Harmonisation des données de différentes plate formes (WhatsApp, Messenger, etc.)
- Normalisation ?
- Anonymisation des données
  - Techniques TAL : détection des entités nommés
  - Grammaires locales

Méthode Phase 3 : Développement du modèle informatique

- Apprentissage machine pour distinguer l'historique d'un patient et d'un participant sain

Méthode Phase 4 : Évaluation

- Auprès de 200 participants âgés sans problème de mémoire :
  - Collecte des historiques des conversations téléphoniques
  - Le modèle informatique estime quels participants montrent une même courbe d'évolution que les patients Alzheimer.
  - Prise de sang pour déterminer s'ils sont porteur de l'Apolipoprotéine E4 (ApoE4), qui donne un risque augmenté de développer la maladie d'Alzheimer. (Hauser and O Ryan 2013)

- Tests statistique :
  Les porteurs de l'ApoE4 ont-il une courbe d'évolution langagière similaire aux patients atteints de la maladie d'Alzheimer ?

Conclusion

Résumé

- Le langage est une faculté cognitive
- Elle s'appuie fortement sur d'autres facultés cognitives comme la mémoire
- Le langage a des propriétés statistiques importantes
- On peut donc étudier le langage grâce aux modèles statistiques
- L'inspiration du TAL est important pour travailler sur des corpus
  - Robustesse
  - Performance
- Plusieurs façons pour explorer Cognition & Linguistique Computationnelle
  - Métrique de coût cognitif
  - Détection de pathologies (Alzheimer)

Recherche d'un Doctorant sur la MA et le TAL

Candidatures bienvenues !

Bibliography I

- Barrett, Maria, Željko Agić, and Anders Søgaard (2015). "The Dundee Treebank". In: *The 14th International Workshop on Treebanks and Linguistic Theories (TLT 14)*.
- Becker, James T, François Boiler, Oscar L Lopez, Judith Saxton, and Karen L McGonigle (1994). "The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis". In: *Archives of Neurology* 51.6, pp. 585–594.
- Brysbaert, Marc and Françoise Vitu (1998). "Word skipping: Implications for theories of eye movement control in reading". In: *Eye guidance in reading and scene perception*. Elsevier, pp. 125–147.
- Chen, Jun, Ji Zhu, and Jieping Ye (2019). "An Attention-Based Hybrid Network for Automatic Detection of Alzheimer's Disease from Narrative Speech". In: *Proc. Interspeech 2019*, pp. 4085–4089.

Bibliography II

- Ehrlich, Kate and Keith Rayner (1983). "Pronoun assignment and semantic integration during reading: Eye movements and immediacy of processing". In: *Journal of Verbal Learning and Verbal Behavior* 22.1, pp. 75–87.
- Frank, Stefan L (2010). "Uncertainty reduction as a measure of cognitive processing effort". In: *Proceedings of the 2010 workshop on cognitive modeling and computational linguistics*. Association for Computational Linguistics, pp. 81–89.
- Fraser, Kathleen C, Jed A Meltzer, and Frank Rudzicz (2016). "Linguistic features identify Alzheimer's disease in narrative speech". In: *Journal of Alzheimer's Disease* 49.2, pp. 407–422.
- Goodglass, Harold and Edith Kaplan (1983). "Boston diagnostic examination for aphasia". In: *Philadelphia: Lea and Febiger*.

Bibliography III

- Hale, John (2001). "A probabilistic Earley parser as a psycholinguistic model". In: *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*. Association for Computational Linguistics, pp. 1–8.
- Hanseeuw, Bernard (2019). "Alzheimer's disease: a clinical research perspective". In:
- Hauser, Paul S and Robert O Ryan (2013). "Impact of apolipoprotein E on Alzheimer's disease". In: *Current Alzheimer Research* 10.8, pp. 809–817.
- Kennedy, Alan, Robin Hill, and Joël Pynte (2003). "The dundee corpus". In: *Proceedings of the 12th European conference on eye movement*.

Bibliography IV

- Lee, Kenton, Luheng He, Mike Lewis, and Luke Zettlemoyer (2017). "End-to-end Neural Coreference Resolution". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 188–197.
- MacWhinney, Brian, Davida Fromm, Margaret Forbes, and Audrey Holland (2011). "AphasiaBank: Methods for studying discourse". In: *Aphasiology* 25.11, pp. 1286–1307.
- McDade, Eric and Randall J Bateman (2017). "Stop Alzheimer's before it starts". In: *Nature News* 547.7662, p. 153.
- Rayner, Keith (1998). "Eye movements in reading and information processing: 20 years of research.". In: *Psychological bulletin* 124.3, p. 372.

Bibliography V


Rowe, Christopher C, Kathryn A Ellis, Miroslava Rimajova, Pierrick Bourgeat, Kerryn E Pike, Gareth Jones, Jurgen Fripp, Henri Tochon-Danguy, Laurence Morandeau, Graeme O'Keefe, et al. (2010). "Amyloid imaging results from the Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging". In: *Neurobiology of aging* 31.8, pp. 1275–1283.


Seminck, Olga and Pascal Amsili (2018). "A Gold Anaphora Annotation Layer on an Eye Movement Corpus". In: *11th edition of the Language Resources and Evaluation Conference*.


Shannon, Claude E and Warren Weaver (1949). *The mathematical theory of communication (Urbana, IL*.


Swaab, Dick Frans (2014). *We are our brains: a neurobiography of the brain, from the womb to Alzheimer's*. Spiegel & Grau.

Bibliography VI


Van Deemter, Kees and Rodger Kibble (2000). "On coreferring: Coreference in MUC and related annotation schemes". In: *Computational linguistics* 26.4, pp. 629–637.