Colloque HUMANUM — 18 octobre 2018

Classification de textes anglais L2 par niveau de compétence langagière

ou

Le problème est-il bien posé ?

ou

L'opacité du deep learning, où nous mène-t-elle ?

Yannis Haralambous
(IMT Atlantique & UMR CNRS 6285 Lab-STICC)
yannis.haralambous@imt-atlantique.fr

Ἐν ἀρχῇ ἦν… À l'origine était… une compétition

Accueil | Appel | Comités | Compétition | Inscription | Programme

Invités | Venue | Soutiens | Contact

Compétition : My Tailor is rich! Prédiction du niveau en anglais à partir de production écrite d'apprenants

Mise à Jour

- 18.05.2018 : Un résumé d'une ou deux pages décrivant la méthodologie que vous avez suivie doit accompagner la soumission.

Final Leaderboard

Rang	Equipe	Score
1.	Balikasg	3.4856
2.	Terislepacom	7.2862
3.	ICSI	8.5970
4.	Caoutchouc	9.4537
5.	ACNK	10.4350
6.	reciTAL team	11.2112
7.	TAU	12.8222
8.	Capitaine-Ad-Hoc	14.1696
9.	Chamlia	17.5161
10.	Haralambous + Lenca Team	17.9774

Plan de l'exposé

1. Le problème, tel qu'il a été posé (planches 4–18)
2. Interrogations (planches 19–19)
3. Solutions proposées (planches 20–28)
4. ~~Humanités~~ Machinalités numériques (planches 29–34)
5. Ce qu'on aurait pu faire (planches 40–42)

Le corpus

- Le corpus EFCAMDAT de l'université de Cambridge est une trace des «essais» écrits par les apprenants L2 de l'anglais de l'école en ligne *Englishtown*.
- Un apprenant qui suit tout le programme, écrira 128 «essais».
- Le programme est subdivise en 16 niveaux :

Table 1: Englishtown skill levels in relation (indicative) to common standards

Englishtown	1-3	4-6	7-9	10-12	13-15	16
Cambridge Esol	-	KET	PET	FCE	CAE	-
IELTS	-	<3	4-5	5-6	6-7	>7
TOEFL iBT	-	-	57-86	87-109	110-120	-
TOEIC Listening & Reading	120-220	225-545	550-780	785-940	945	-
TOEIC Speaking & Writing	40-70	80-110	120-140	150-190	200	-
CEFR	A1	A2	B1	B2	C1	C2

(tiré de Huang *et al.*, 2017)

Le corpus

- Quelques exemples d'essais :

Table 2: Examples of essay topics at various levels. Level and unit number are separated by a colon.

ID	Essay topic	ID	Essay topic
1:1	Introducing yourself by email	7:1	Giving instructions to play a game
1:3	Writing an online profile	8:2	Reviewing a song for a website
2:1	Describing your favourite day	9:7	Writing an apology email
2:6	Telling someone what you're doing	11:1	Writing a movie review
2:8	Describing your family's eating habits	12:1	Turning down an invitation
3:1	Replying to a new penpal	13:4	Giving advice about budgeting
4:1	Writing about what you do	15:1	Covering a news story
6:4	Writing a resume	16:8	Researching a legendary creature

(tiré de Huang *et al.*, 2017)

Le corpus

- 1,2 millions de textes (Pennington *et al.*, 2014)
- 175 mille apprenants (de 198 pays différents)
- le challenge : apprendre la classification à partir de 27 310 textes et tester l'algorithme appris sur un ensemble de test de 13 665 textes.

Le référentiel de compétences linguistiques

- Niveau A : utilisateur élémentaire (= scolarité obligatoire), lui-même subdivisé en niveau introductif ou de découverte (A1) et intermédiaire ou usuel (A2).
- Niveau B : utilisateur indépendant (= lycée), subdivisé en niveau seuil (B1) et avancé ou indépendant (B2). Il correspond à une «compétence opérationnelle limitée» (Wilkins) ou une «réponse appropriée dans des situations courantes» (Trim).
- Niveau C : utilisateur expérimenté, subdivisé en C1 (autonome) et C2 (maîtrise).

Plus précisément

PROFICIENT USER	C2	Can understand with ease virtually everything heard or read. Can summarise information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation. Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in more complex situations.
	C1	Can understand a wide range of demanding, longer texts, and recognise implicit meaning. Can express him/herself fluently and spontaneously without much obvious searching for expressions. Can use language flexibly and effectively for social, academic and professional purposes. Can produce clear, well-structured, detailed text on complex subjects, showing controlled use of organisational patterns, connectors and cohesive devices.
INDEPENDENT USER	B2	Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.
	B1	Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst travelling in an area where the language is spoken. Can produce simple connected text on topics which are familiar or of personal interest. Can describe experiences and events, dreams, hopes & ambitions and briefly give reasons and explanations for opinions and plans.
BASIC USER	A2	Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment). Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters. Can describe in simple terms aspects of his/her background, immediate environment and matters in areas of immediate need.
	A1	Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. Can introduce him/herself and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has. Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help.

Le corpus

- Exemple de texte «C2» :

  The three robots mentionned in the web-based magazine are or will be very useful in the future.The first one, the Escort Humanod Robot will have emotional intelligence and be able to have facial mouvement to express sadness or happiness. It could help a lot of lonely people. The second one is already improving the health and medecine area. Florence (the robot's name) has already spread some services like checking monitoring and emptying trash, but would be develop to become a great help to visit quarantine patient suffering from contagious illnesses and be able to monitor care.And for the last type of robots, the industrial one. It can be used to clear up natural or man made disaster as well as for leisure, and leisure is also part of life.

- Exemple de texte «A1» :

  hi, my name's melanie. I'm twenty years old and i live in France. i would like to learn english because I would to go in usa.

Indicateurs fournis

- sentences : nombre de phrases ;
- nombre de mots ;
- nombre de lettres ;
- nombre de syllabes ;
- nombre de signes de ponctuation ;
- nombre moyen de mots par phrase ;
- taille moyenne des mots en nombre de caractères ;
- nombre moyen de syllabes par mots ;
- nombre de phrases par mot ;
- formule mettant en jeu le nombre de mots d'une seule syllabe ;
- variante mettant en jeu également en jeu le nombre mots divisé par le nombre de phrases ;
- 66 indices de lisibilité.

Indices de lisibilité fournis 1/6

- ARI (Automated Readability Index). La formule de calcul met en jeu des coefficients du nombre de mots divisés par le nombre de syllabes et le nombre de prépositions ;
- Bormuth, qui donne une estimation du niveau scolaire nécessaire pour comprendre un texte. Il est fondé sur la liste des 3 000 mots les plus fréquents en anglais (liste Dale-Chall) ;
- indice de lisibilité proportionnel aux nombre de lettres et au nombre de phrases (tous les 100 mots) ;
- indice de lisibilité (1995), qui reflète le degré de familiarité du lexique utilisé et qui repose sur la liste des 3 000 mots les plus fréquents en anglais (liste Dale-Chall) ;
- deux formules de lisibilité qui reposent sur le nombre de caractères utilisés (espaces compris) ;

Indices de lisibilité fournis 2/6

- indice de lisibilité qui prend en compte des valeurs proportionnelles au nombre de mots, au nombre de caractères ;
- mesure des degrés de lisibilité (Degrees of Reading Power) à partir de l'indice de Bormuth ;
- (Easy Listening Formula) : le nombre de mots polysyllabiques divisé par le nombre de phrases ;
- indice proche de Flesch, mais où le nombre de mots monosyllabiques tous les 100 mots simplifie la prise en compte du nombre de syllabes tous les 100 mots ;
- métrique sensible à la langue utilisées. Ce sont les valeurs pour l'anglais qui ont été utilisées. On retranche à 206 835 des valeurs proportionnelles au nombre de mots divisés. L'indice est compris entre 100 (texte facile à comprendre) et 0 (texte très difficile) ;

Indices de lisibilité fournis 3/6

- métrique développée pour les besoins de l'armée américaine pour proposer une conversion de la difficulté d'un texte en niveau scolaire nécessaire pour le lire ;
- indice de lisibilité proposé dans les années cinquante. Il est censé représenter le nombre d'années d'études nécessaires à la compréhension d'un texte à la première lecture. Il prend en compte la proportion de mots de trois syllabes ou plus ;
- indice de lisibilité collectivement (FORCAST = Patrick FORd, John CAylor and Thomas STicht) mis au point avec des conscrits du Vietnam, qui repose sur la longueur des mots. On retranche à 20 le dixième des mots monosyllabiques (sur une fenêtre de 150 mots) ;
- caractéristique stylistique proposée par W. Fucks (produit du nombre de caractères divisé par le nombre de mots et du nombre de mots divisé par le nombre de phrases) ;

Indices de lisibilité fournis 4/6

- indice qui prend en compte le nombre de mots de trois syllabes ou plus, le nombre de mots et le nombre de phrases ;
- proposé au départ pour l'analyse du suédois, cet indice prend en compte la proportion de mots de 7 lettres ou plus. Les textes qui ont un indice inférieur à 25 sont censés être faciles à lire, sont «normaux» autour de 40 et considérés comme difficiles au-delà de 50 ;
- indices proposés dans les années quatre-vingts pour l'allemand (Neue Wiener Sachtextformeln), qui prennent en compte, dans des proportions variables, les mots de trois syllabes ou plus et les mots de six lettres ou plus ;
- adaptation pour l'anglais de l'indice LIX. Il s'agit du nombre de mots de six lettres ou plus divisé par le nombre de phrases ;

Indices de lisibilité fournis 5/6

- Simple Measure of Gobbledygook (SMOG). Indice de lisibilité fondé sur la racine carré du nombre de mots polysyllabiques, calculés au début, au milieu, début et fin de texte ;
- indice de lisibilité fondé sur le nombre de mots du texte qui ne figurent pas dans la liste de mots de références de Spache ;
- indice de lisibilité des médias mis au point en 2006 qui divise le nombe de syllabes des trois premières phrases par dix ;
- indices de lisibilité prenant en compte la proportion de prépositions (Traenkle.Bailer.TB1) et de conjonctions (Traenkle.Bailer.TB2) ;
- indice réputé ne pas être sensible à la langue analysée, il met en jeu le logarithme du nombre de mots divisé par le nombre de phrases ;
- indice de lisibilité proposé dans les années cinquante, qui met en jeu le nombre de deux syllabes ou plus ;

Indices de lisibilité fournis 6/6

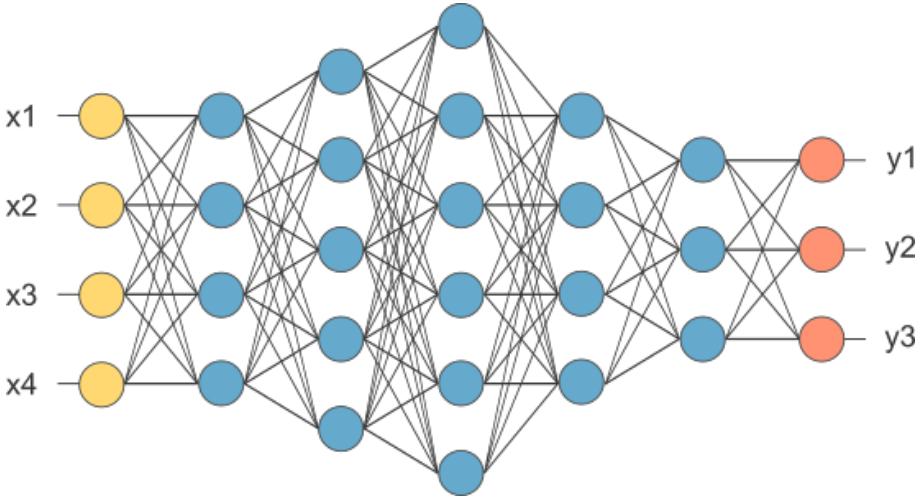
- algorithme proposé par Carroll pour corriger le TTR ;
- indice de diversité lexicale fondé sur la probabilité de retrouver un mot dans une fenêtre de 42 mots ;
- indice C de Herdan ;
- indices de complexité lexicale proposés en 1972 qui mettent en jeu les logarithmes des types et des tokens ;
- moyenne mobile du type to token ratio. (Moving Average of TTR) calculée par le biais d'une fenêtre mobile. Si le texte comporte moins de 400 mots, le MATTR ne peut pas être calculé et renvoie NA.

(liste tirée de Huang *et al.*, 2017)

Ces indicateurs sont basés sur le principe que *ratatouille* (4 syl.) est plus difficile que *hubris* (2 syl.), etc. («lexicométrie»)

Cf. Conquet & Richaudeau (1973).

Outil recommandé pour le challenge



Les réseaux de neurones (cf . Chollet (2018) pour une bonne introduction).

Comment les réseaux de neurones gèrent-ils le texte ?

- Les phrases sont découpées en «mots» (ni lexèmes, ni morphèmes).
- Ceux-ci forment une structure ensembliste (le «sac de mots»).
- Ils deviennent des coordonnées dans un espace vectoriel de grande dimension. On appelle cela des *word embeddings*. (Exemple : le *GloVe* développé par Stanford en 2015, 400 kmots, 300 dimensions, 822 Mo.)
- Dans cet espace :
  1. la distance euclidienne est proportionnelle à la «similarité sémantique» et
  2. certaines directions privilégiées représentent «plus ou moins» des correspondances linguistiques (genre, nombre, cas, substantivation, etc.).

Interrogations

Interrogations spécifiques au corpus :

- Sous quelles conditions les textes ont-ils été écrits ?
- Peut-on généraliser alors que le public est très hétérogène ?
- Quel rapport entre qualité de production et compétences ?
- Est-il pertinent d'utiliser des mesures de lisibilité ?

Interrogations générales :

- Quid de la syntaxe ?
- À quoi cela nous servirait-il de mesurer la similarité des textes entre eux ?
- À quoi sert un outil opaque ? Que nous apprend-il sur la langue ?

Quelques solutions apportées : coquilles

- Les niveaux *graphématique* et *orthographique*.
- On compte les «coquilles» = mots qui n'apparaissent pas dans le dictionnaire de formes,
- ni dans Wikipédia ou dans WordNet, en tant qu'entités nommées.
- On affecte un poids inférieur aux «coquilles» qui donnent des mots existants quand on remplace une lettre par celle dont la touche se trouve à une distance de 1 de la lettre correcte (= «faute de frappe»).
- On affecte un poids supérieur aux «coquilles» qui se répètent dans le même texte.

La ponctuation

- Le niveau *orthotypographique*.
- On répertorie les espaces manquants ou superflus (en particulier ceux qui relèvent d'autres conventions nationales).
- On vérifie la capitalisation (en particulier du pronom personnel *I*).
- Les erreurs d'élision (**I'am *I'ts *I'have*, etc.).
- Il ne s'agit pas de détecter une «connaissance des règles de ponctuation» mais plutôt l'état de stress ou de sérénité du scripteur.

Conjugaison / verbes à particule

- Le niveau *morphologique*.
- On compte les erreurs d'accord (**two friend*), de conjugaison (**I sayed*), la confusion entre participe et gérondif (**I seen him*), les erreurs de particule (**I go at the theater*).
- Parmi les verbes à particule (*verbal phrases*), on établit la distribution probabiliste des correspondances verbe/particule (par exemple *add on* = «ajouter» est plus rare que *add up* = «additionner»), et on les pondère par l'inverse de leur probabilité d'occurrence.
- Un poids négatif est attribué aux verbes à particule erronés.

Cf. Swan & Smith (2001) : les erreurs fréquentes par nationalité !

Simple English Wikipedia

- Le niveau *lexical*.
- Les mots que l'on retrouve dans le *Simple English Wikipedia* sont pondérés positivement vis-à-vis de ceux que l'on retrouve dans le Wikipedia standard.

Simple English Wikipedia :

Cluj-Napoca is the third biggest city in Romania, and is the capital city of Cluj County, in the north-western part of Transylvania. Bucharest is about 330 kilometers away from Cluj-Napoca. About 330,000 people live in the city.

English Wikipedia :

Cluj-Napoca is the fourth *most populous* city in Romania, and the *seat* of Cluj County in the northwestern part of the country. *Geographically*, it is *roughly equidistant* from Bucharest, Budapest and Belgrade. As of 2011, 324,576 *inhabitants* lived *within the city limits*, *marking a slight increase from the figure recorded at the 2002 census*.

Indicateurs de Kyle-Crossley

Propriétés psycholinguistiques des mots :

- Concrétude : *apple* > *infinity*.
- Familiarité : *breakfast* > *egress* (= «sortie»).
- *Imageability* (facilité de se forger une image) : *beach* > *philology*.
- *Meaningfulness* (taille du voisinage pour la mesure de similarité sémantique) : *beautiful* > *adze* (= «aissette»).
- Âge d'acquisition : *apple* = 2,11 < *calculus* = 6,89.

Cf. (Gilhooly & Logie, 1980) et (Kyle & Crossley, 2014).

Kyle `www.linguisticanalysistools.org` propose un outil pour calculer ces indicateurs : *TAALES* (*Tool for the Automatic Analysis of Lexical Sophistication*).

Tentatives de se servir de la syntaxe

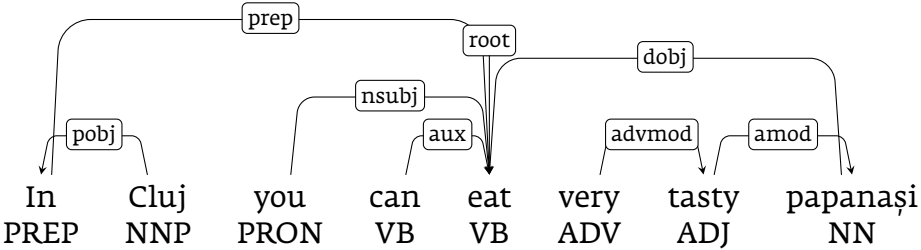
- Les réseaux de neurones utilisent le sac de mots (ou des sacs de *n*-grammes).
- La syntaxe est un indicateur potentiel de la qualité de production :

  **Would you confirming me that you've well been receive this email ?* (A1) («Pourriez-vous me confirmer que vous avez bien reçu ce mail ?»)

- Comment gérer la structure syntaxique indépendamment des lexèmes ?

Arbres syntaxiques par dépendances

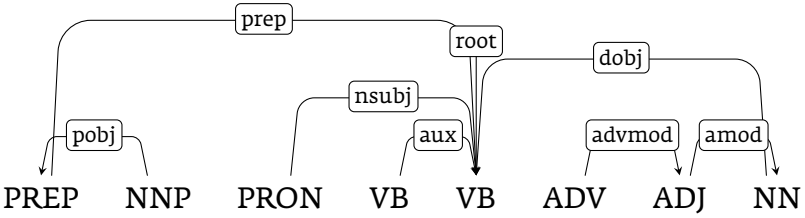
On peut représenter la structure syntaxique d'une phrase sous forme d'un arbre enraciné et étiquetté :



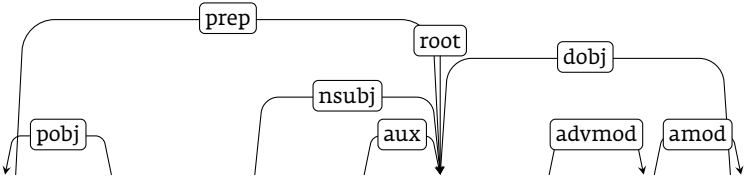
Les étiquettes des sommets sont les parties du discours, celles des arêtes sont les natures de dépendance.

Arbres syntaxiques par dépendances

Nous avons calculé les arbres sans les occurrences de mots :

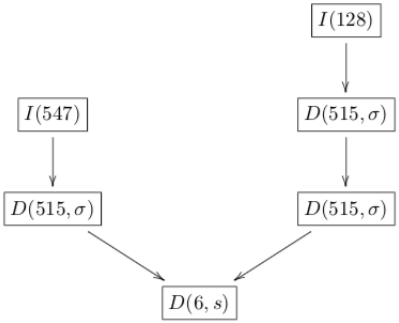


ainsi que sans les parties du discours :



Apprentissage d'arbres syntaxiques

- Nous avons créé un plongement de graphes en utilisant l'approche *graph2vec* (Narayanan *et al.*, 2017).
- Un tel plongement permet de calculer la similarité (structurelle et sémantique) des graphes en tant que distance euclidienne dans un espace vectoriel de grande dimension (dans notre cas, 128).
- Nous avons utilisé un réseau de neurones *feed-forward* avec double entrée (données textuelles et indicateurs d'un côté, arbres syntaxiques de l'autre), comme suit :



~~Humanités~~ Machinalités numériques

- Nous avons réussi à prédire le niveau européen de compétence langagière avec une précision de 89,86 %.

~~Humanités~~ Machinalités numériques

- Nous avons réussi à prédire le niveau européen de compétence langagière avec une précision de 89,86 %.
- Pour le corpus EFCAMDAT, il s'entend.

~~Humanités~~ Machinalités numériques

- Nous avons réussi à prédire le niveau européen de compétence langagière avec une précision de 89,86 %.
- Pour le corpus EFCAMDAT, il s'entend.
- Qu'avons-nous appris sur l'usage de la langue par les apprenants ? Sur l'acquisition des différentes structures ?

~~Humanités~~ Machinalités numériques

- Nous avons réussi à prédire le niveau européen de compétence langagière avec une précision de 89,86 %.
- Pour le corpus EFCAMDAT, il s'entend.
- Qu'avons-nous appris sur l'usage de la langue par les apprenants ? Sur l'acquisition des différentes structures ?
- Rien.

~~Humanités~~ Machinalités numériques

- Nous avons réussi à prédire le niveau européen de compétence langagière avec une précision de 89,86 %.
- Pour le corpus EFCAMDAT, il s'entend.
- Qu'avons-nous appris sur l'usage de la langue par les apprenants ? Sur l'acquisition des différentes structures ?
- Rien.
- À quoi cela sert-il d'apprendre/comprendre/connaître si la machine peut effectuer efficacement des tâches *sans que l'on ait à lui transmettre quelque logique que ce soit* ?

~~Humanités~~ Machinalités numériques

- Le *deep learning* est-il en train de déshumaniser l'informatique ?

~~Humanités~~ Machinalités numériques

- Le *deep learning* est-il en train de déshumaniser l'informatique ?
- (Zhang, Zhao & Le Cun, 2016) arrivent (à l'aide d'un réseau convolutionnel de 27 couches) à classifier des textes en partant des caractères, *sans utiliser quelque ressource linguistique que ce soit*. La linguistique et les linguistes sont-ils inutiles ?

~~Humanités~~ Machinalités numériques

- Le *deep learning* est-il en train de déshumaniser l'informatique ?
- (Zhang, Zhao & Le Cun, 2016) arrivent (à l'aide d'un réseau convolutionnel de 27 couches) à classifier des textes en partant des caractères, *sans utiliser quelque ressource linguistique que ce soit*. La linguistique et les linguistes sont-ils inutiles ?
- Quel est l'avenir du traitement automatique de la langue ?

~~Humanités~~ Machinalités numériques

- Le *deep learning* est-il en train de déshumaniser l'informatique ?
- (Zhang, Zhao & Le Cun, 2016) arrivent (à l'aide d'un réseau convolutionnel de 27 couches) à classifier des textes en partant des caractères, *sans utiliser quelque ressource linguistique que ce soit*. La linguistique et les linguistes sont-ils inutiles ?
- Quel est l'avenir du traitement automatique de la langue ?
- Deux scénarios :

~~Humanités~~ Machinalités numériques

- Le *deep learning* est-il en train de déshumaniser l'informatique ?
- (Zhang, Zhao & Le Cun, 2016) arrivent (à l'aide d'un réseau convolutionnel de 27 couches) à classifier des textes en partant des caractères, *sans utiliser quelque ressource linguistique que ce soit*. La linguistique et les linguistes sont-ils inutiles ?
- Quel est l'avenir du traitement automatique de la langue ?

- Deux scénarios :
- (pessimiste) Sans apport linguistique, des machines de plus en plus puissantes vont effectuer des tâches de plus en plus complexes *par simple apprentissage automatique*. La linguistique perdra progressivement son utilité.

~~Humanités~~ Machinalités numériques

- Le *deep learning* est-il en train de déshumaniser l'informatique ?
- (Zhang, Zhao & Le Cun, 2016) arrivent (à l'aide d'un réseau convolutionnel de 27 couches) à classifier des textes en partant des caractères, *sans utiliser quelque ressource linguistique que ce soit*. La linguistique et les linguistes sont-ils inutiles ?
- Quel est l'avenir du traitement automatique de la langue ?

- Deux scénarios :
- (pessimiste) Sans apport linguistique, des machines de plus en plus puissantes vont effectuer des tâches de plus en plus complexes *par simple apprentissage automatique*. La linguistique perdra progressivement son utilité.
- (optimiste) L'étude des structures plus complexes de la langue apportera des améliorations aux systèmes de *deep learning* — *plus la linguistique progressera, plus ces systèmes s'amélioreront.*

Ce qu'on aurait pu faire...

- ...mais qu'on n'a pas fait, par manque de données ou de temps.
- Nous possédons tous un modèle L (compétence ou connaissance ?) d'une langue donnée (phonologie/graphématique, lexique, morphologie, syntaxe, sémantique, pragmatique, etc.).
- L'apprentissage d'une langue consiste à faire évoluer

$$\begin{tikzcd}
\cdots \arrow[r] & L_{i-1} \arrow[r] & L_i \arrow[r] & L_{i+1} \arrow[r] & \cdots \\
 & & L_{\text{standard}} \arrow[ul] \arrow[u] \arrow[ur] & &
\end{tikzcd}$$

- Si on avait les identifiants d'utilisateurs on aurait pu étudier l'évolution des modèles d'une production à l'autre.

Ce qu'on aurait pu faire…

- on aurait pu exploiter l'origine de l'apprenant pour identifier les difficultés spécifiques, comparer le modèle de la langue L1 de l'apprenant avec son L_i et L_S ;
- on aurait pu faire une inférence grammaticale sur chaque production pour trouver les dérivations syntaxiques utilisées et les comparer (distance entre grammaires formelles à définir) avec celles de L_S ;
- on aurait pu classifier les essais par thème (= exercice donné), construire une ontologie du thème basée sur L_S et comparer les sous-graphes des productions avec ceux de L_S ;
- mais on ne l'a pas fait. Comme disait Hippocrate déjà 25 siècles avant nous…

Hippocrate disait…

Ὁ βίος βραχύς,	Ars longa,
ἡ δὲ τέχνη μακρά,	vita brevis,
ὁ δὲ καιρὸς ὀξύς,	occasio praeceps,
ἡ δὲ πεῖρα σφαλερή,	experimentum periculosum,
ἡ δὲ κρίσις χαλεπή.	iudicium difficile.

La vie est courte,
l'art long,
l'opportunité précipitée,
l'expérimentation faillible,
le jugement difficile.

Références bibliographiques I

- CAp2018 (2018), *Appel à participation à la compétition «my taylor is rich» de CAp 2018*, http://cap2018.litislab.fr/competition_annexes_FR.pdf.
- Chollet, F. (2018), *Deep Learning with Python*, Manning, Shelter Island.
- Conquet, A. & Richaudeau, F. (1973) «Cinq méthodes de mesure de la lisibilité», *Communications & Langages* 17, p. 5–16.
- Gilhooly, K. J. & Logie, R. H. (1980) «Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words», *Behavior Research Methods & Instrumentation* 12, p. 395–427.
- Haralambous, Y. & Lenca, P. (2018), *Response to the "My Tailor is Rich" Challenge*, http://perso.telecom-bretagne.eu/yannisharalambous/ressources/.
- Huang, Y. *et al.* (2017), *The EF Cambridge Open Language Database*, https://corpus.mml.cam.ac.uk/efcamdat2/public_html/EFCamDat-Intro_release2.pdf

Références bibliographiques II

- Kyle, K. & Crossley, S. A. (2014) «Automatically Assessing Lexical Sophistication: Indices, Tools, Findings, and Application», *tesol* 49, p. 757–786.
- Narayanan, A. *et al.* (2017), *graph2vec: Learning Distributed Representations of Graphs*, arXiv : 1707.05005.
- Pennington, J. *et al.* (2014), «GloVe: Global Vectors for Word Representation» in *Proceedings of EMNLP 2014*, p. 1532–1543.
- Swan, M. & Smith, B. (2001), *Learner English: a teacher's guide to interference and other problems*, Cambridge University Press.
- Zhang, X., Zhao, J. & Le Cun, Y. (2016), *Character-level Convolutional Networks for Text Classification*, arXiv : 1509.01626.

Home, sweet home

Les transparents de cette présentation sont disponibles ici :

http://perso.telecom-bretagne.eu/yannisharalambous/ressources/